Table of Contents

COMPOSITIONS AND METHODS FOR DIRECTED GENE ASSEMBLY

The present application claims priority under 35 U.S.C. § 119(e) to U.S. provisional application Serial No. 60/222,139, filed July 31, 2000, the entire contents of which is incorporated herein by reference in its entirety.

1. INTRODUCTION

The present invention is directed to methods and compositions for use of homologous recombination for directed evolution, gene reassembly, and directed mutagenesis. One aspect of the present invention relates to methods and compositions for use of bacterial conjugative transfer and homologous recombination for directed evolution, gene reassembly, and directed mutagenesis.

2. BACKGROUND OF THE INVENTION

Evolution can be viewed as an algorithm wherein a sequence gives rise to variants and a selection performed on that derivative pool allows for the survival of progeny with an incremental enhancement of the selected trait (Daniel C. Dennett, *Darwin's Dangerous Idea*, Touchstone, New York, NY 1995). Iterative cycles of the process drive the production of increasingly refined embodiments of the selected trait. In popular models of natural evolution the global "fitness" of the organism is the driving selective force. Beginning at the dawn of civilization man has intervened in the process to exert selections on potential food corps and animals, not for the fitness of the organism, but rather for utility to his kind. This is a "directed evolution".

In recombinant DNA technologies, individual genes can be isolated and expressed in foreign host organisms allowing the controlled production of specific gene products. This ability forms the basis of the biotechnology industry, with applications in medicine, agriculture and various chemical industries (see, *e.g.*, Evens and Witcher, 1993, Ther. Drug Monit. 15(6):514-20; Steve Prentis, Biotechnology: a new industrial revolution, G. Braziller, NY, NY 1984; Symposium on Biotechnology for Fuels and Chemicals, Totowa, N.J.: Humana Press, 1997). With recombinant DNA technologies, and the isolation of individual genes directed evolution procedures can be applied to these isolated genes. The term "directed evolution", as commonly used, applies to efforts made to improve the characteristics of a gene product with a particular commercial end in mind (Marrs *et al.*, 1999, Curr. Opin. Microbiol 2(3):241-5), although in some instances the term has been applied to groups of genes defining a pathway (Wackett, 1998, Ann. N Y Acad. Sci. 864:142-52).

The first efforts to accomplish this involved the application of various mutagenesis procedures that introduce changes at single, or at times several, residues of the coding sequence (Kuchner and Arnold, 1997, Trends Biotechnol. 12:523-30). Such efforts have reported some success, albeit, limited. The number of potential changes to be explored is immense, vastly exceeding an experimenter's ability to produce and analyze them. It is clear that most changes are detrimental while only rare alterations yield enhancements in desired trait.

More recently, specialized PCR technologies have been applied to the problem of directed evolution (Stemmer, 1994, Proc. Natl. Acad. Sci. 91:10747-51). The most popular version, primerless PCR or so-called sexual PCR, allows for the re-assortment, or "shuffling," of closely related sequences. Briefly, a set of related gene sequences are fragmented, denatured, allowed to reanneal, and PCR extension is then performed through a number of cycles to reconstruct unit length genes. This process produces novel sequences that are complex permutations of the substrates. This process has proven to produce genes with significantly varied characteristics, and in many instances phenotypes dramatically improved for selected properties (*e.g.*, Chang *et al.*, 1999, Nat. Biotechnol. 8:793-7). In a set of experiments with a related family of β-lactamases, mutagenesis was compared directly with shuffling. The shuffling procedure proven to dramatically enhance resistance to a novel β-lactam (500-fold) where only modest improvements (8-fold) were noted in with mutagenesis alone (Crameri *et al.*, 1998, Nature 391:288-91). Both mutagenesis and re-assortment sample an array of potential variants. When sampling re-assorted variants, the set of sampled sequences contains variants that are composed of sequence stretches that have themselves been "pre-selected," over evolutionary time scales, for function. This is in contrast to the sequences derived from mutagenesis where the combinations are likely to be encountered for the first time without "pre-selection." The hypothesized "pre"-selection aspect of this re-assortment procedure may allow for the apparently more productive nature of the so-called shuffling strategy.

Although "gene shuffling" has had some success and can be credited with popularizing the notion that cloned genes can be tailored to provide more useful variants through directed evolution procedures, it has clear limitations that make alternative strategies desirable. For example, one major shortcoming of "shuffling," or more precisely, random complex permutation sampling, is that information about a particular member of a combinatorial set only becomes accessible when the exact identity of that member is revealed. When complex permutations are sampled randomly, as in so-called gene shuffling, any information about the context of the sample is lost until its identity is revealed, following sequence determination. Furthermore, random permutation sampling through primerless PCR is a process that requires all subsequent iterations to repeat the

enzymatic steps of the process: DNA isolation, DNA fragmentation, PCR reconstruction, and product cloning. A faster and more cost-effective procedure would be desirable.

Plasmid-based recombination has previously been used as an approach for producing novel genes (Piotukh *et al.*, 1992, Molekulyarnaya Biologiya 26(4) part 2:601-604) used homologous recombination to construct hybrid metalloproteinases. This approach used direct repeat recombination, a process requiring only a single crossover event. Such recombination can produce novel genetic arrangements, but each round of iteration requires re-cloning of the sequences targeted for the recombination process, and reagents used for one event cannot reused or archived for subsequent procedures. Although a highly efficient process, this type of recombination does not lend itself to combinatorial reassortments or multiuse libraries.

Citation of a reference herein shall not be construed as an admission that such is prior art to the present invention.

3. SUMMARY OF THE INVENTION

The present invention provides methods and compositions for directed gene assembly ("DGA") that generate pluralities of divergent DNA molecules that can be used for functional analysis and directed evolution of genes ("target genes") in a laboratory setting. In these methods, a vector-borne donor molecule provides sequences that recombine with sequences of a vector-borne target molecule through homologous recombination to direct the assembly of divergent DNA molecules. In the present invention the directed assembly is achieved independently of the phenotypic characteristics encoded by the target sequences. Rather, selection is based on marker sequences physically linked to the target sequences. The resultant variant target molecules make possible a variety of subsequent selections or screens that may be executed on a diverse plurality of the recombinant products. Such subsequent screens can often be executed in a second host organism (other than the host in which the recombination event is selected) where prior enrichment for the recombinant product is required to make the process tractable.

Such bimolecular homologous recombination events allow for substrates of the process to be used repeatedly in iterative combinatorial exchanges. With respect to such iterations, the present invention involves directed, rather than random, iterative exchanges based on information obtained by analysis of the variants obtained in the previous iteration(s). Since the substrates are vectors that replicate in a host cell, *e.g.*, a bacterium, they can be archived. For example, information about the potential function of substrates of the process may be deliberately sought by directing exchanges with sequences encoding structurally or empirically characterized target proteins.

The combination of archival and repeated uses results in historical information that leads to databases. For example, the present invention can be employed to create collections of protein structural domains that may be employed in directed evolution procedures that take into account increasing information about protein structure.

In its simplest form, the present invention involves a vector-based system that works through direct pair-wise exchanges between a donor and a target. This is in marked contrast to exchanges that can be catalyzed in primerless PCR strategies where multiple parents are made to participate in an exchange resulting in complex permutation sampling. In general, complex permutation sampling is not desirable because it only provides useful information from those members of a library for which full sequence information is determined, and as a consequence is not a powerful strategy for guiding subsequent iterative rounds. Unlike the PCR strategy, the DGA donor/target strategy of the invention proceeds in a logical and directed manner based on a systematic search where the iterative rounds involve mixing cells, *e.g.*, bacteria, without new rounds of molecular biology procedures.

Advantages of the methods of the present invention are exemplified in Section 6, *infra*. Generation of variants of bacterial subtilisins, which are serine proteases that cleave polypeptides, using DGA produced a >95% yield of variants with functional protease activity (see, *e.g.*, Section 6.6.2). This result can be contrasted with results reported for PCR-based shuffling of subtilisin sequences (Ness *et al.*, 1999, Nature Biotechnology 17: 893-896). In the PCR-based shuffling experiments, only 6% of the resultant products showed protease activity. Thus, DGA methodology, which produces functional variants, significantly reduces the burden of labor-intensive assays required to screen against the 94% inactive products from the PCR procedure.

The donor/target selection described herein is based on the placement of a negative selection sequence into a position in the target sequence where the directed substitution is desired. The process is designed to take advantage of the *in vivo* biological process of homologous recombination. Three kinds of reagents are required for this process: (1) a donor DNA, (2) a target DNA and (3) a negative selection insert in the target DNA in the region where DNA segment replacement is desired. In one embodiment, the product of the homologous recombination is selected for directly, *i.e.*, in a one-step process. In a more preferred embodiment, a two-step procedure is used to select for the product of homologous recombination, which entails selection of an intermediate state in the process followed by selection of the product of homologous recombination. In such an embodiment, the intermediate state is one in which the target cell contains both the donor vector and the target vector. Without wishing to be bound by any theory or mechanism, it is believed that this intermediate state more particularly involves an intermediate of the

homologous recombination process referred to as a co-integrant. In the latter embodiment, a fourth element is required, namely a positively selectable sequence in the donor DNA to allow for selection of the intermediate state.

The invention encompasses, first, a method for generating a population of variant sequence modules in cells, *e.g.*, bacterial cells, said method comprising: (a) transferring a donor vector into a target cell which is capable of homologous recombination, wherein (i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first donor DNA sequence and a second donor DNA sequence; and (ii) said target cell comprises a target vector comprising a target recombination module comprising, in the following order from 5' to 3': a first target DNA sequence; a negatively selectable marker; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence; and (b) selecting for a population of target cells which do not contain the negatively selectable marker, so that a population of variant sequence modules in cells, in particular, the target cells is generated. Generally, selecting for target cells that do not contain the negatively selectable marker is accomplished by subjecting the cells to conditions that do not allow growth of donor cells or of target cells that still contain the negatively selectable marker (*i.e.*, have not undergone recombination with the donor vector resulting in loss of the negatively selectable marker). To ensure loss of donor cells, for example, a selectable marker (*e.g.*, a tetracycline resistance-encoding element) can be included in the chromosomal background of the target cell, but be absent from the donor cell. Imposing appropriate selective pressure (*e.g.*, inclusion of tetracycline) results in selected loss of donor cells. In a variation of this method, the target recombination module is present in the target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

In another embodiment, the invention provides a method for generating a population of a variant sequence modules in cells, *e.g.*, bacterial cells, said method comprising:(a) transferring a donor vector into a target bacterial cell which is capable of homologous recombination, wherein (i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first non-functional fragment of a positively selectable marker; a first donor DNA sequence; and a second donor DNA sequence; (ii) said target cell comprises a target vector comprising a target recombination module comprising, in the following order from 5' to 3': a second non-functional fragment of the positively selectable marker; a first target DNA sequence; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA

sequence, and said second donor DNA sequence is homologous to said second target DNA sequence, and recombination between said first non-functional fragment of the selectable marker and said second non-functional fragment of the selectable marker results in a functional selectable marker; and (b) selecting for a population of target cells which contain a functional positively selectable marker, so that a population of a variant sequence modules in the cells is generated. In a variation of this method, the target recombination module is present in the target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

The cells undergoing DGA, *i.e.*, target cells into which the donor vector has been transferred, are subjected to conditions that allow homologous recombination to take place. Conditions that allow homologous recombination to occur merely refer to standard growth or maintenance conditions for the particular cells being used in the particular instance.

In a preferred embodiment, the donor vector and target vector of the foregoing methods are present in bacterial cells. In one embodiment of the method, the bacterial cell is an *E. coli* cell. In other embodiments, the bacterial cell is a naturally transformable cell such as *Acinetobacter calcoaceticus*, *Haemophilus influenzae*, or *Neisseria meningitidis*. In another preferred embodiment, the donor vector and the target vector are present in a bacterial cell, and said transferring is by conjugative transfer of at least the donor recombination module of the donor vector from the donor cell to the target cell. In other embodiments, the donor vector is transformed into the target cell or is transferred into the target cell via a phage particle.

In another preferred embodiment, the donor vector further comprises a positively selectable marker. Where the donor vector further comprises a positively selectable marker, the methods of the present invention preferably further entail, between step (a) and step (b): (a') selecting for a population of target cells, *e.g.*, bacterial cells, with the donor vector, by selecting for the presence of the positively selectable marker in the donor vector.

In one embodiment, these methods further comprise the step of: (c) selecting said population of target cells which do not contain the negatively selectable marker for a desired phenotype. In another embodiment, the invention provides a method for optimizing a phenotype comprising the above-mentioned method, further comprising: the step of (d) repeating steps (a) - (c), wherein the target recombination module used in step (d) is derived from a target cell selected in step (c), and said selection is based on information obtained from the analysis of the variant sequence modules obtained in step (c).

In another embodiment, the donor vector further comprises a third donor sequence, located 3' to the first donor sequence and 5' to the second donor sequence. In another embodiment, the target recombination module of step (d) is identical to the target recombination module of step (a). In another embodiment, the target recombinant module of step (d) is different from the target recombination module of step (a). In yet another embodiment, the methods further comprise, prior to step (a), the step of mutagenizing the donor DNA vector. In one embodiment, the step of mutagenizing the donor vector is carried out *in vitro*. In another embodiment, the step of mutagenizing the donor molecule is carried out *in vivo*.

In another embodiment, the negatively selectable marker comprises a conditionally lethal sequence and selecting the recombinant comprises selecting against said conditionally lethal sequence. In yet another embodiment, the negatively selectable marker of the target recombination module is a polar insert sequence which prevents expression of a downstream reporter gene, such that deletion of said polar insert results in expression of the reporter gene, and the step of selecting for a population of target cells which do not contain the negatively selectable marker comprises detecting or selecting for expression of said reporter gene. In various embodiments, the polar insert is a Tn5 or a Tn10 sequence.

In certain embodiments, the negatively selectable marker can be selected against on the basis of its physical properties. Such selection is referred to herein as "molecular selection." In one such embodiment, the negatively selectable marker in the target recombination module comprises a unique restriction endonuclease recognition site, and selection for a recombinant variant comprises selecting against molecules with the restriction endonuclease recognition site. In another such embodiment, the negatively selectable marker is selected against on the basis of its size, said selection comprising amplifying DNA from cells to identify and isolate sequences comprising recombinant target modules that have lost the negative selection insert.

In various embodiments of the present invention, there is at least 75%, at least 80%, more preferably at least 85%, yet more preferably at least 90%, and most preferably at least 95% sequence identity between the first donor DNA sequence and the first target DNA sequence and between the second donor DNA sequence and the second target sequence.

In a preferred embodiment of the invention, the donor vector is a suicide vector.

The invention further provides kits suitable for directed assembly of a target DNA molecule. These kits comprise donor vectors, donor cells, target vectors and/or target cells of the invention.

In one embodiment, such a kit comprises in one or more containers: a) a donor vector comprising a donor recombination module comprising, in the following order from 5' to 3': a first donor DNA sequence and a second donor DNA sequence, and b) a target cell which is capable of homologous recombination, said cell comprising a double-stranded DNA target vector useful for directed assembly of a target DNA molecule of interest, said target vector comprising a target recombination module comprising, in the following order from 5' to 3': a first target DNA sequence; a negatively selectable marker; and a second target DNA sequence, such that said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence.

In another embodiment, such a kit comprises, in one or more containers: a) a donor vector, comprising a donor recombination module comprising, in the following order from 5' to 3': a first non-functional fragment of a positively selectable marker, a first donor DNA sequence, and a second donor DNA sequence; b) a target cell comprising a target vector comprising, in the following order from 5' to 3': a second non-functional fragment of the positively selectable marker; a first target DNA sequence; and a second target DNA sequence, wherein them said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence, and recombination between said first non-functional fragment of the selectable marker and said second non-functional fragment of the selectable marker results in a functional selectable marker. In one embodiment, the donor vector is present within a cell, *i.e.*, a donor cell.

In one kit embodiment, the donor vector further comprises a third donor sequence, located 3' to the first donor sequence and 5' to the second donor sequence. In another kit embodiment, the donor vector further comprises a positively selectable marker. In a preferred embodiment, the cells of the kit are bacterial cells, preferably *E. coli* cells or naturally transformable bacterial cells.

The invention further provides libraries suitable for the practice of directed gene assembly. Such libraries can be donor or vector libraries and can comprise a plurality of any of the donor or target vectors of the invention, including vectors comprising variant target sequences that have been produced via DGA. Such libraries can also comprise variant target gene or target gene sequences produced via DGA that no longer contain intervening selectable markers and encode variant target gene products, including optimized variant target gene products. Libraries can also comprise a plurality of archived sequences or modules, optionally present within cells.

The invention further encompasses databases of archived modules. An archived module, as used herein, refers to a donor DNA sequence or target DNA sequence,

whether or not the donor or target sequence has undergone DNA or phenotype optimization, where the sequence comprising the archived module is known or has been demonstrated to encode a protein segment or domain that provides a particular function and has been stored and cataloged (archived).

The present invention still further provides a computer readable medium having a database recorded thereon in computer readable form, wherein said database comprises one or more module profiles and wherein each module profile describes a phenotype in a DGA assay, and wherein each module profile is associated with a particular vector in a particular target cell.

4. BRIEF DESCRIPTION OF THE FIGURES

FIG. 1. *Features of the Donor Vector.* FIG. 1 shows a donor vector (Section 5.1.1) comprising a recombination module (referred to as "bc"), described in Section 5.1.1.1; a selectable marker, described in Section 5.1.1.4; an origin of replication, which is preferably conditional and compatible with the target vector, described in Section 5.1.1.3; and, optionally, where the donor vector is to be transferred into the target cell by means of conjugation, conjugative transfer sequences as described in Section 5.1.1.2.

FIG. 2. *Features of the Target Vector.* The left panel FIG. 2 shows a target vector Section 5.1.2) comprising a target recombination module ("ABCDE"), described in Section 5.1.2.1; a selectable marker and an origin of replication for propagation of the target vector in the target cell (Section 5.1.2.2), and, optionally, an additional selectable "shuttle" origin of replication and selectable marker that can be used to propagate the vector in a different cell (Section 5.1.2.2). The right panel shows a target vector as in the left panel, further comprising negatively selectable marker galK, which is the galactokinase gene under the control of the galactose operator ("galOP"). This negatively selectable marker (see Section 5.1.2.1.1) imparts galactose sensitivity on target cells with a *galE*⁻ genotype that comprise the target vector.

FIG. 3. *Method for Selecting Recombinant Product: Selection Against Non-Recombinants.* FIG. 3 shows how the product of a DGA event can be selected for (see Section 5.2) by selecting against a negatively selectable marker ("xyz") present in the target recombination module ("ABC"). A recombination event between the donor recombination module ("abc") in which the strand crossover sites flank the negatively selectable marker results in the generation of a variant target module ("ABc") lacking the negatively selectable marker. The negatively selectable marker can be selected against (see Section 5.2.1) to identify recombinant variant target vectors with variant target modules.

FIG. 4. *Method for Selecting Recombinant Product: Elimination of Polar Sequence.* FIG. 4 shows how the product of a DGA event can be selected for (see Section

5.2) by selecting against a polar sequence such as Tn10 present in the target recombination module ("ABC"). The target vector further comprises a promoter sequence on the 5' side of the target recombination module and a reporter gene ("wxyz") placed 3' of the target recombination module (see Section 5.1.2.1.1). The polar sequence inhibits expression of the reporter gene. A recombination event between the donor recombination module ("abc") in which the strand crossover sites flank the polar sequence results in the generation of a variant target module ("ABc") lacking the polar sequence. In the absence of the polar sequence, the reporter gene ("wxyz") can be transcribed and selected for (see Section 5.2.1) to identify recombinant variant target vectors with variant target modules.

FIG. 5. *Method for Selecting Recombinant Product: Reconstruction of Flanking Selectable Marker.* FIG. 5 shows how the product of a DGA event can be selected for (see Section 5.2) by reconstruction of a reporter gene ("wxyz"), as described in Section 5.1.2.1.2. The target vector comprises a non functional fragment of the reporter gene ("wxy") and the donor vector comprises a second, complementary non-functional fragment ("xyz") of the reporter gene. A recombination event between the donor recombination module ("ABC") and the target recombination module ("abc') results in the generation of a variant target module ("ABc") and a functional reporter gene ("wxyz"), which can be expressed and selected for (see Section 5.2.1) to identify recombinant variant target vectors with variant target modules.

FIG. 6. *Directed mutagenesis.* FIG. 6 shows a DGA process essentially as described in FIG. 3, but where the donor is mutagenized (as described in Section 5.2) prior to DGA, to produce a mutagenized donor recombination module ("a*b*c*"). DGA using a mutagenized donor results in the variant target module "Abc*".

FIG. 7. *Gene Family Re-Assortment.* FIG. 7 shows how starting with a target recombination module ("ABCD") with two negatively selectable markers ("gal" and "sac") allows for 2 successive rounds of DGA, thereby generating greater diversity. In the example of FIG. 7, for each target recombination module ("ABCD"), the first round of DGA utilizes two related donor modules ("ab" and "a'b'") homologous to the "AB" portions of the target recombination module and the second recombination step utilizes another two related donor modules ("cd" and "c'd'") homologous to the "CD" portions of the target recombination module. Gene family re-assortment is described in Section 5.2.5.

FIG. 8. *Identification of structural motifs.* FIG. 8 describes how a novel protein ("AbCD") can be generated by DGA starting with a target recombination module ("ABCD") comprising a negatively selectable marker in "B" and a donor recombination module "b". The activity of "AbCD" can be compared with the activity of "ABCD" to determine whether "b" can functionally substitute for "B". This information can be used to generate additional variants. See Section 5.2.5.

FIG. 9. *Insertional acquisition and substitution.* FIG. 9 shows how DGA can be utilized to replace sequences "DE" in the target recombination module ("ABCDEF", where a negatively selectable marker ("xyz") is inserted into "D") with non-homologous sequences "δϵ". This is achieved by subjecting the target recombination module to DGA with a donor recombination module ("Cδϵ F") in which the non-homologous sequences are flanked by homologous sequences ("C" and "F"). See Section 5.3.

FIG. 10. *Selection for Segregation of Donor Vector.* FIG. 10 show a DGA process, essentially as described in FIG. 3, utilizing a "suicide" donor vector. A suicide donor vector has an origin of replication compatible with the cell in which the donor vector is propagated (*e.g.*, a donor cell) but is incompatible with the target cell. This DGA configuration allows the elimination of recombined donor vectors following DGA.

FIG. 11. *Sequence isolation.* FIG. 11 shows how DGA can be utilized to isolate novel homologous sequences to a target sequence of choice from a nucleic acid library. Nucleic acid sequences from a library (*e.g.*, "c", "b", "a") are inserted as donor recombination modules into a donor vector. Using the negative selection method described in FIG. 3, only recombination events that result in deletion of the negatively selectable marker ("xyz") from the target recombination module ("ABC" comprising the negative selection marker in "C") are identified. In the example shown in FIG. 11, the donor recombination module "c" will recombine with the target recombination module to generate "ABc", thereby identifying "c" as a homologous sequence to "C". This method is described in Section 5.2.5.

FIG. 12. *Creation of extracted libraries.* FIG. 12 shows an "extracted donor library", in which donors producing products with desired properties are set aside to produce the extracted library, which is a specialized library containing modules or sequences of similar or related function. See Section5.3.

FIG. 13. *Iterative cycling of Product to Target.* FIG. 13 shows how a target recombination sequence ("ABCDEF") can be activated by insertion of a negatively selectable marker ("gal") by DGA in which the "activating" donor recombination module ("BCDE") comprises the negatively selectable marker in "B". After one round of DGA between the activated target ("ABCDEF" comprising the "gal" marker in "B") with a diversity donor ("ab"), which produces the variant "AbCDEF", the new product can be activated with another activating donor ("BCDE") with a negatively selectable marker in a different position (in "D"). A second round of DGA with a second diversity donor ("cd") produced yet another variant product ("AbCdEF"). This process can be repeated to produce large numbers of substrates for future rounds of DGA.

FIG. 14. *Schematic of two step co-integrant formation and resolution.* FIG. 14 shows the generation of a co-integrant, which is an intermediate of homologous

recombination, which comprises target vector sequences (including an AMP selection cassette) and donor vector sequences (including a gentamycin resistance cassette). Selection against the negatively selectable marker ("xyz") in the target recombination module ("ABC") will select for recombination products of DGA. For further details, see Section 5.2.2.

FIG. 15. *Schematic of pGPG plasmid series creation and features.* For additional details on the construction of the pGPG plasmid series, see Section 6.1.1.

FIG. 16. *Sequence of 3A13 and 5A20 sequences in pGPG.* For a description of these plasmids, see Section 6.1.2.

FIG. 17. *Sequence of complete* lichenformis *(5A36) and* subtilis *(3A1) subtilisins in target vector.* For details, see Section 6.2.1.

FIG. 18. *Schematic representation of selectable / negative selection inserts.* For details, see section 6.2.3.

FIG. 19. *Schematic representation of reduced target vectors.* For details of these vectors, see Section 6.2.5.

FIG. 20. *Diagram of Gal-Spec and Kan-Suc inserts in target vectors.* For details of these vectors, see Section 6.2.4.

FIG. 21. *Diagram showing principles of restriction nuclease-based selection against unrecombined target and donor vectors.* Such methods are described in Section 5.1.2.1.1.

FIG. 22. *Diagram showing principles of PCR size-based molecular selection against unrecombined target and donor vectors.* Such methods are described in Section 5.1.2.1.1.

FIG. 23. *Schematic and data describing use of DGA to place inserts unto target vector.* FIG. 23 show how DGA (with the molecular restriction nuclease-based selection) can be used to insert donor sequences into a stretch of homologous target DNA, as described in Section 6.5.3.

FIG. 24. *Table of oligonucleotides used in this study.*

5. DETAILED DESCRIPTION OF THE INVENTION

Described herein are methods and compositions for directed gene assembly ("DGA"). The DGA system can iteratively be utilized until an optimized sequence for a desired trait has "evolved". First, a target sequence of interest is subjected to a systematic process that results in variation within the target. Variation is preferably generated by conjugative transfer of donor sequences into the target cell followed by homologous recombination between a donor sequence and the target sequence, as discussed in detail herein. The present invention also encompasses the use of methods other than conjugation

to transfer the donor vector into the target cell, including but not limited to transformation or phage-mediated transfer. Second, the resulting sequence variants can be subjected to a selection process in which the sequences are selected or screened for exhibition of a desired trait. One or more iterations of the DGA process can be utilized to further optimize a desired trait. The starting material for each subsequent iteration is based on information obtained via analysis of variants obtained in the prior iteration(s). For example, sequence information obtained can indicate what domain or domains of the target sequence should be targeted for further sequence variation. Thus, rather than producing purely random sequence variants from a variant obtained in one round of DGA, the present method involves iterative DGA to systematically generate truly directed variants. Such DGA cycles can be reiterated as many times as necessary until sufficient optimization of the sequence of interest for the desired trait is attained.

The methods for DGA described herein utilize classical molecular and genetic techniques. In a preferred embodiment, DGA exploits the techniques of bacterial conjugation and homologous recombination. The DGA system is based on a collection of donor vectors and target vectors, and donor vector and target vector libraries. The target vectors are constructed and transformed into host strains, thereby creating target cellular, *e.g.*, bacterial cell, populations. The donor vectors can, for example, be in the form of transformable plasmids or phage genomes. In a more preferred embodiment, donor vectors are constructed and transformed into host strains, thereby creating donor cellular populations. In one embodiment, donor and target cell populations are bacterial cell populations. In another embodiment, the donor and target cell populations are bacterial cell populations that are designed to be capable of bacterial conjugation with each other, such that, upon mixing of the donor and target cell populations, bacterial conjugation allows delivery of donor vector sequences from the donor cell to the target cell. Once the donor vector sequences that include the donor recombination module are in a target cell which expresses homologous recombination activity, homologous recombination results in rearrangement of target DNA sequences, due to regions of sequence homology between the donor and target gene sequences.

As used herein, two sequences are "homologous" if they share a region of sequence identity, optionally interrupted by one or more mis-matched base pairs, such that they are capable of homologous recombinational exchange with each other. In a preferred embodiment, two homologous double-stranded sequences are completely identical. In another embodiment, the extent of homology is interrupted by not more than 1 mismatched base pair every approximately 10 base pairs of identical nucleotides. In a preferred embodiment, the extent of homology is a continuous stretch of at least 30, 40, 50, 60, 70, 80 90 or 100 base pairs of identical nucleotides. In various embodiments, the extent of

homology between homologous sequences is a continuous stretch of at least 6, 8, 10, 15, 20, 25, 30, 35, 40, 50, 60, 75 or 100 base pairs of identical nucleotides. In an alternative embodiment, a stretch of identical nucleotides can be interrupted by 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 non-identical nucleotides per 100 identical nucleotides. In yet other embodiments, the extent of sequence identity between donor sequences and target sequences (*i.e.*, each pair of first and second sequences) is at least 70%, more preferably at least 75%, more preferably at least 80%, more preferably at least 85%, yet most preferably at least 90% or 95% identity. In certain specific embodiments, the extent of sequence identity between donor and target sequences is at least 92%, 94%, 96%, 98% or 99%. Homologous sequences may be interrupted by one or more non-identical residues, provided they are still efficient substrates for homologous recombination.

The use of homologous recombination to promote rearrangements, particularly when coupled with bacterial conjugation, allows successive iterations of "evolution cycles" without requiring new rounds of *in vitro* molecular biological manipulations. Thus, this system provides faster and more cost effective methods, as compared to other methods for directed evolution, such as gene shuffling approaches.

Described below, are compositions and methods relating to DGA systems. In particular, Section 5.1 describes compositions suitable for practicing DGA, including donor vectors and libraries, target vectors and libraries, and cells carrying such vectors. Section 5.2, below, describes the DGA methods, including methods for the generation and selection of sequence variants, methods for optimization of a desired trait, and methods for reiteration of the DGA process. Finally, Sections 5.3, 5.4 and 5.5, below, describe archived collections of libraries and databases.

5.1 COMPOSITIONS SUITABLE FOR USE IN DIRECTED GENE ASSEMBLY

In this section, compositions suitable for practicing DGA, including donor vectors and libraries, target vectors and libraries, and cells carrying such vectors are described in detail.

5.1.1 THE DONOR VECTOR

The invention encompasses donor vectors and donor vector libraries. A summary of the basic characteristics of the donor vector is presented in FIG. 1. Briefly, the donor vector comprises a donor recombination module, optionally a conjugative transfer element, and standard sequences required for maintenance and propagation of the donor vector in the cell, such as an origin of replication and a selectable marker. The donor vector

can optionally further comprise a multiple cloning site and/or an additional selectable marker, in particular a positively selectable marker.

Preferably, the donor vector contains only a minimum amount of vector sequence homologous to other standard vectors, if any at all. Such a feature limits the amount of unwanted homologous recombination between donor and target vectors. Nonetheless, appropriate selection schemes can readily be devised to select against such rare, extra-recombination module recombination events. It is noted that the homology referred to herein refers to homology outside the donor and target recombination modules, and, in appropriate embodiments, outside the first and second non-functional selectable marker fragments.

The features of the donor vector are described in detail hereinbelow. The DNA vectors described herein may be constructed using standard methods known in the art (see Sambrook *et al.*, 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York; Ausubel, *et al.*, 1989-1999, Current Protocols in Molecular Biology, Green Publishing Associates and Wiley Interscience, N.Y., both of which are incorporated herein by reference in their entirety). For example, synthetic or recombinant DNA technology may be used. Oligonucleotides may be synthesized using any method known in the art (*e.g.*, standard phosphoramidite chemistry on an Applied Biosystems 392/394 DNA synthesizer). Further, reagents for synthesis may be obtained from any one of many commercial suppliers. Finally, it is noted that a donor vector can be constructed or derived from what is referred to herein as a "pre-donor" vector or plasmid molecule. Such a pre-donor molecule comprises the donor vector features described herein, including a multiple cloning site, but lacks a complete donor recombination module. In one embodiment, the pre-donor molecule contains a first and second donor DNA sequence within the multiple cloning site, but lacks a selectable marker between the two donor DNA sequences. Subsections of Section 6.1, below, describe the construction of pre-donor molecules.

5.1.1.1 THE DONOR RECOMBINATION MODULE

The donor vector contains at least two regions of sequence homology to a target vector: a first donor DNA sequence which is homologous to a first target DNA sequence; and a second donor DNA sequence which is homologous to a second target DNA sequence, so that homologous recombination can occur between donor and target vectors in a cell which capable of supporting homologous recombination (see *e.g.*, Doherty *et al.*, 1983, J. Mol. Biol. 167: 539-60; and Laban and Cohen, 1981, Mol. Gen. 184: 200-7 for general discussion of homologous recombination). These regions of sequence homology reside within "recombination modules" on the respective vectors, with a donor

recombination module on the donor vector, and a target recombination module on the target vector.

The donor recombination module comprises, in the following order from 5' to 3': a first donor DNA sequence; optionally, a third donor DNA sequence; and a second donor DNA sequence. The first and second donor DNA sequences are homologous to sequences in the target DNA and are designed so that homologous recombination between these sequences and the target DNA sequences will occur and result in sequence exchange. Upon homologous recombination between homologous sequences of the donor and target vectors, sequences residing between the regions of homology are exchanged, creating a new product comprising a sequence variant of the target sequence. This product comprises a variant product module.

The optional third donor sequence is not homologous to sequences in the target DNA and is preferably a negatively selectable sequence (see Section 5.1.2.1.1, *infra*), which can be present either alone or in conjunction with a positively selectable marker (preferably different from the positively selectable marker or markers present elsewhere on the donor vector), for example, present as part of a selectable marker cassette. Such cassettes include, but are not limited to, Gal-Spec and Kan-Suc cassettes, as described in Section 6, below.

Homologous recombination in the target cell results in strand exchange between donor and target DNA sequences. To the extent that target DNA sequences are part of or correspond to target gene sequences, in general, it is preferred that the donor vector comprises donor DNA sequences homologous to a portion of the target gene smaller than the entire target gene. Such a situation represents another way the extent of the region exchanged in the recombination process can be directed to a particular region of the target gene. To avoid recombination events outside the target gene in the target vector, the donor vector is preferably designed so that the only target gene homology lies within the recombination module. Thus, the donor vector should generally not share sequence homology of 10 or more contiguous base pairs with the target vector outside the target recombination module.

DNA sequences for use with these vectors and methods may come any source, including, but not limited to, prokaryotic, archaebacterial or eukaryotic DNA sequences, or from viral, phage or synthetic origins. For example, nucleic acid sequences may be obtained from the following sources: human, porcine, bovine, feline, avian, equine, canine, insect (*e.g., Drosophila*), invertebrate (*e.g., C. elegans*), plant, microbial (*e.g.*, thermophilic bacteria) *etc.* In one embodiment, the DNA donor sequences are derived from characterized cloned DNA sequences and libraries of sequences. In another embodiment, the source DNA is produced synthetically, for example, by synthesizing oligonucleotides.

Stretches of random oligonucleotides can be embedded into the homologies used to direct recombination in the DGA strategy of the present invention. The random nucleotides can be present as continuous stretches or be interspersed with regions of fixed homologies; so long as homolgous recombination between the target and donor sequences can still occur. The use of random synthetic sequences within the context of the DGA approach to directed evolution has broad application to all the potential applications of directed evolution, extending well beyond those of typical random peptide libraries which most typically address specific binding interactions (see, *e.g.*, Brown, 2000, Curr. Opin. Chem. Bio. 14: 16-21). In another embodiment, sequences are generated by mutagenesis, and cloned into donor vectors that can be used in DGA.

Such DNA sequences may be obtained and used to construct donor vectors suitable for use in the DGA system, as described above, by standard procedures known in the art, such as, for example, standard molecular biology techniques, such as PCR and molecular cloning, *etc.* (*see, e.g.*, Sambrook *et al.*, 1989, *supra*; Ausubel, *et al.*, *supra*; Glover (ed.), 1985, DNA Cloning: A Practical Approach, M.L. Press, Ltd., Oxford, U.K. Vol. I, II). Libraries of donor vectors may be archived, for multiple use with different target vectors (see Sections 5.3 and 5.4 hereinbelow).

5.1.1.2 CONJUGATIVE TRANSFER SEQUENCES

In embodiments where the donor vector is transferred to target cells by means of conjugative transfer, the donor vector comprises sequences that direct the conjugative transfer of donor DNA sequences from the donor cell to the target cell using the process of bacterial conjugation process. During conjugation, a physical bridge is formed between two bacterial cells which allows the exchange of plasmid DNA (see, *e.g.*, Wallets and Wilkins, 1984, Microbiol. Rev. 48: 24-41).

Systems for bacterial conjugation, and the genes and sequences required therefor, are known in both gram-negative bacteria, including *E. coli* (see Nunez *et al.*, 1997, Mol. Microbiol. 24: 1157-68; Wallets and Skurray,1987, Cellular and Mol. Biol. (Ed. F.C. Neidhardt) pp. 1110-1133); Wallets and Skurray, 1982, Ann. Rev. Genet. 14:41-76) and gram-positive bacteria (Firth *et al.*, 1999, Mol. Microbiol. 31:1598-600). Such sequences can be inserted into donor vector, which can confer conjugative transferability to donor plasmid. In one embodiment, conjugative transfer functions are provided *in trans* from gene sequences present on the bacterial chromosome, thereby preventing transfer of the donor vector in the target cell (Metcalf *et al.* 1994, Gene 138:1-7). Donor cells designed in this way also produce more copies of the vector on which the conjugative transfer sequence resides. In a preferred embodiment, sequences from the conjugative plasmid R6K is used (for review see Filutowicz and Rakowski, 1998, Gene 223:195-204). The donor

vector contains a minimal *cis*-acting R6K sequence, while *trans*-acting conjugation genes required for recognition, transfer, and structural functions are provided by sequences on the chromosome of the target cell. Minimal lambda phages designed for convenient cloning are well known to those trained in the art of molecular biology (see Miller 1992, *supra*). Derivatives with conditional lethal mutations that require propagation in an amber suppressor host provide a convenient gene delivery system as infection of such phage into a bacterium without an amber suppressor produces in no infection and simply results in the delivery of DNA to a host strain, *i.e.*, a target cell.

5.1.1.3 ORIGIN OF REPLICATION

The donor vector requires an origin of replication, which is needed for propagation of the vector. In this respect, donor vectors must be designed to be compatible with other plasmids in the donor cell, as well as target cell vectors.

For cloning and propagation in *E. coli*, any *E. coli* origin of replication may be used, examples of which are well-known in the art (see Miller, 1992, A Short Course in Bacterial Genetics, Cold Spring Harbor Laboratory Press, NY, and references therein). Non-limiting examples of readily available plasmid origins of replication are ColE1-derived origins of replication (Bolivar *et al.*, 1977, Gene 2:95-113; see Sambrook *et al.*, 1989, *supra*), p15A origins present on plasmids such as pACYC184 (Chang and Cohen, 1978, J. Bacteriol. 134:1141-56; see also Miller, 1992, *supra,* p.10.4-10.11), and pSC101 origin are all well known in the art.

For example, in one embodiment, a high-copy replicating plasmid is used, such as a plasmid containing a ColE1-derived origin of replication, examples of which are well known in the art. One example is an origin from pUC19 and its derivatives (Yanisch-Perron *et al.*, 1985, Gene 33:103-119), which have convenient cloning sites for insertion of foreign genes. An example of a medium-copy plasmid with a ColE1-derived origin of replication is pBR322 (Bolivar *et al.*, 1977, Gene 2:95-113; see Sambrook *et al.*, 1989, *supra*).

In one embodiment, a donor plasmid having a p15A origin of replication is used, to be compatible with a target plasmid having a ColE1-derived origin of replication. One example of a plasmid having a p15 origin of replication is pACYC184, one of the pACYC100 series of plasmids, which exist at 10-12 copies per cell (Chang and Cohen, 1978, J. Bacteriol. 134:1141-56; see also Miller, 1992, p. 10.4-10.11). In another embodiment, another ColE1 compatible plasmid, pSC101 origin, such as pSC101, which exists at approximately 5 copies per cell, may be used. Both pACYC and pSC101 plasmid vectors have convenient cloning sites and can co-exist in the same cell as pBR and pUC plasmids.

Other suitable plasmid origins of replication include lambda or phage P1 replicon-based plasmids, for example the Lorist series (Gibson *et al.*, 1987, Gene 53: 283-286). In another embodiment, synthetic origins of replication may be used. In another embodiment, non-plasmid vectors may also be used. For example, λ vectors, such as λgt11 (Huynh *et al.*, 1984, in "DNA Cloning Techniques: A Practical Approach," Vol I, D. Glover, ed., pp 49-78, IRL Press, Oxford), or the T7 or SP6 phage systems (Studier *et al.*, 1990, Methods Enzymol. 185:60-89) can be used. Such viral systems would not require conjugation for delivery of DNA sequences.

In yet another embodiment, the origin of replication of a donor vector and/or a target vector is compatible with replication in a *Salmonella* species, most preferably *Salmonella typhimurium*. For examples of origins of replications compatible with *Salmonella*, see, *e.g.,* Miller, J.H., 1992, A Short Course in Bacterial Genetics, Cold Spring Harbor Laboratory Press, NY; Neidhardt, F.C., ed., 1987, *Escherichia coli and Salmonella typhimurium*, American Society for Microbiology, Washington, D.C.

The positively selectable sequence can be present at any position of the donor DNA vector as long as it does not interfere with vector functions (for example replication in donor cells, conjugative transfer, if utilized, other selectable markers present on the vector). Among the cells that can be utilized in conjunction with the vectors and methods described herein are naturally transformable bacterial cell such as *Acinetobacter calcoaceticus* (ATCC No. 33305). An exemplary origin of replication that can be utilized in vectors intended to be present in *A. calcoaceticus* is the origin of replication preset in the cryptic plasmid pWH1277 described in Hunger *et al.*, 1990, Gene 87:45-51.

In a preferred embodiment, the origin is a conditional origin of replication, that is, is one that is dependent on transacting replication functions that are not present in the target cell. For a discussion of such a transacting factor see Kruger *et al.*, 2001, J Mol. Biol. 306:945-55. Constructed in this way, the donor vector cannot replicate in the target cell, thereby facilitating its loss after it is transferred into the target cell.

5.1.1.4 SELECTABLE MARKERS

To maintain the donor vector in the cell, the vector typically contains a selectable marker. Any selectable marker known in the art can be used. Donor vectors must be compatible with vectors of the target cell, which requires the choice of a selectable marker different than, and compatible with any selectable markers expressed in the target cell. Any gene that conveys a readily identifiable or selectable phenotypic change, such as resistance to an antibiotic effective in *E. coli*, can be used. Preferably, the selectable marker is an antibiotic resistance gene, such as the kanamycin resistance gene from TN903 (Friedrich and Soriano, 1991, Genes. Dev. 5:1513-1523), or genes that confer resistance to

other aminoglycosides (including but not limited to dihydrostreptomycin, gentamycin, neomycin, paromycin and streptomycin), the β-lactamase gene from IS1, that confers resistance to penicillins (including but not limited to ampicillin, carbenicillin, methicillin, penicillin N, penicillin O and penicillin V). Other selectable genes sequences including, but not limited to gene sequences encoding polypeptides which confer zeocin resistance (Hegedus *et al.* 1998, Gene 207:241-249). Other antibiotics that can be utilized are genes that confer resistance to amphenicols, such as chloramphenicol, for example, the coding sequence for chloramphenicol acetyltransferase (CAT) can be utilized (Eikmanns *et al.* 1991, Gene 102:93-98). As will be appreciated by one skilled in the art, other non-antibiotic methods to select for maintenance of the plasmid may also be used, such as, for example a variety of auxotrophic markers (see Sambrook *et al.*, 1989, *supra*; Ausubel *et al.*, *supra*), which can be selected by adding or subtracting a particular nutrient from the growth media.

5.1.2 THE TARGET VECTOR

In addition to the donor vector, the invention also encompasses target vectors and target vector libraries. A summary of the basic characteristics of the target vector is presented in FIG. 2. Briefly, the target vector comprises a target recombination module, preferably, sequences that allow transcription of sequences within the first and/or second target DNA sequences of the target recombination module, and standard sequences required for maintenance and propagation of the vector in the cell,.*e.g.*, an origin of replication and a selectable marker.

Preferably, the target vector contains only a minimum amount of vector sequence homologous to other standard vectors, if any at all. Such a feature limits the amount of unwanted homologous recombination between donor and target vectors. Nonetheless, appropriate selection schemes can readily be devised to select against such rare, extra-recombination module recombination events. It is noted that the homology referred to herein refers to homology outside the donor and target recombination modules, and, in appropriate embodiments, outside the first and second non-functional selectable marker fragments.

In a specific embodiment, a target vector comprises a target recombination module, one or more origins of replication, and one or more selectable markers (*e.g.*, an antibiotic resistance gene). The target recombination module is designed to enable selection for recombinant variant target vectors, and to provide a mechanism for selection against non-recombinant vectors, both donor and targets. To this end, as described in detail below, a recombinant selection system is built into sequences within or immediately flanking the target recombination module.

In a variation of this method, the target vector is present in the target cell integrated into the target cell genome. In such an embodiment, the vector need not contain sequences required for maintenance and propagation of the vector and, therefore, comprises a target recombination module. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

Finally, it is noted that a target vector can be constructed or derived from what is referred to herein as a "pre-target" vector or plasmid molecule. Such a pre-target molecule comprises the target vector features described herein, including a multiple cloning site, but lacks a complete target recombination module. In one embodiment, the pre-target molecule contains a first and second target DNA sequence within the multiple cloning site, but lacks a selectable marker between the two target DNA sequences. Section 6.2.1, below, describe the construction of pre-target molecules.

5.1.2.1 THE TARGET RECOMBINATION MODULE

The target vector comprises a target recombination module comprises, in the following order from 5' to 3': a first target DNA sequence and a second target DNA sequence. The target recombination module further comprises additional sequences to select products of recombination. As discussed in detail below, such sequences can allow selection against non-recombined target vectors using negatively selectable markers, and/or for recombined target molecules using positively selectable markers.

Target sequences from which the variant sequences are generated by the methods of the invention can include any DNA sequence of interest. For example, a target sequence can encode a polypeptide of interest, or a fragment thereof (*e.g.*, a structural or biological domain of the polypeptide of interest). Among the nucleic acid sequences that can be varied according to the methods of the present invention are ones that encode polypeptides that include, but are not limited to polypeptides, or portions thereof, involved in cell proliferation, development, differentiation, signal transduction, enzymatic reactions, either *in vivo or in vitro*. Alternatively, for example, a target sequence can be a regulatory sequence *e.g.*, a sequence that controls, positively or negatively, the temporal and/or spatial, cell or tissue-specific expression, of a coding region to which the regulatory sequence is operably attached. In another embodiment, a target sequence encodes a nucleic acid, *e.g.*, an antisense or ribozyme molecule, that can modulate the expression of a gene or transcript *in trans*.

5.1.2.1.1 NEGATIVELY SELECTABLE MARKERS

In one embodiment, the target vector comprises a target recombination module comprising, in the following order from 5' to 3': a first target DNA sequence; a negatively selectable marker, and a second target DNA sequence. The first and second target DNA sequences are respectively homologous to sequences in the first and second donor DNA sequences, described in Section 5.1.1.1, above, designed so that homologous recombination between donor and target sequences results in sequence exchange.

The negatively selectable marker is included in the target recombination module to facilitate selection for target recombination modules that have successfully undergone homologous recombination. In principle, any negative selection system that allows selection against non-recombined target vector can be used for DGA. Examples of such negatively selectable markers are provided hereinbelow.

In one embodiment, the negatively selectable marker is a sequence that encodes a conditional lethal gene product, whose expression is detrimental to cell growth under a particular set of conditions. Recombination results in the exchange of the negatively selectable marker. Under selective conditions, the lethal function is expressed, and only the recombined products with variant recombination modules will survive. This selection for recombinant products does not depend on the precise nature of the specific recombinant products (FIG. 3). A large number of conditionally lethal sequences are known which can be used in these assays, including, but not limited to, sucrose sensitivity (Lawes and Maloy, 1995, J. Bacteriol. 177:1383-7), and galactose sensitivity (Ahmed, 1984, Gene 28:37-43). Selection against the conditionally lethal marker will enrich for the sub-population of recombinants which can be tested for the desired phenotype.

In another embodiment, the negatively selectable marker is a polar sequence (see FIG. 4). Certain sequences, such as sequences found within the transposon *Tn5* or *Tn10* have the capacity to block the progress of RNA polymerase along a DNA template, resulting termination of transcription. Thus, the presence of these so-called polar sequences can block the expression of downstream genes (Merrick *et al.*, 1978, Mol. Gen. Genet. 165: 103-11).

For the purposes of a negatively selectable marker comprising a polar sequence, it is necessary to construct a target vector comprising: a promoter sequence on the 5' side of the first target DNA sequence, a polar sequence placed in the target recombination module 3' to the first target DNA sequence, and 5' to the second target DNA sequence, and a reporter gene placed downstream from the 3' side of the second target DNA sequence, such that expression of the reporter gene is dependent upon transcription initiated at the promoter sequence and continuing through the recombination module. Thus, the presence of the polar sequence within the module blocks transcription unless homologous

recombination between donor and target sequence results in the removal of the polar sequence and expression of the reporter gene. Selection for the expression of the downstream reporter gene requires the removal of the polar insert.

For the purposes of this selection scheme a "reporter gene" sequence can comprise any gene sequence which expresses or encodes a detectable or positively selectable gene product (preferably a protein). In a preferred embodiment, the activity or presence of such a gene product allows cell growth under selective conditions. A variety of such reporter gene sequences well known to those of skill in the art can be utilized. For example, β-lactamase, which confers resistance to the penicillin family of antibiotics can be used, or sequences which confer resistance to other antibiotics, such as tetracycline, streptomycin, gentamycin, neomycin, kanamycin, hygromycin, or chloramphenicol. Non-antibiotic methods, such as, for example, auxotrophic markers (see Sambrook *et al.*, 1989, *supra*; Ausubel *et al.*, *supra*) may also be used as reporter genes, as will be appreciated by one skilled in the art, other which can be selected on particular growth conditions. Detectable markers suitable as reporter genes include but are not limited to β-galactosidase and green fluorescent protein.

In other embodiments, the negatively selectable marker is any nucleic acid having a sequence that confers certain physical properties that can be the subject of selection. For example, in one embodiment, the negatively selectable marker is a nucleic acid sequence comprising a restriction enzyme recognition site that is unique to the target vector and thus absent from the variant target produced by homologous recombination. The digestion of a mixture of molecules containing the resolved recombinant structure with the enzyme will convert the target vector, but the not the desired recombinant target variant, from a circular to a linear molecule. Circular molecules are more effective at transforming cells than linear molecules, and this difference can be dramatically enhanced by subsequent phosphatase or exonuclease treatment. In this way a property of the insert (as revealed by nuclease treatment) can provide a molecular selection for the recovery of the desired recombinant class.

In another selection method based on the physical properties of a sequence inserted into the target module, the general property of DNA length (without regard to sequence particulars) can also provide a mechanism to select recombinant molecules using PCR. By limiting the extension time of a PCR reaction driven by primers outside the target gene, PCR reaction extension time can be used to size-select the amplification of a desired product class. A thermostable polymerase will proceed at a rate of about 17 bases per second requiring about 90 seconds to complete a 1.5KB segment. Thus, for example, if such a segment had an insert of an additional 4KB, PCR could not amplify the 5.5 KB target with a 90 second extension time. If both the 1.5 KB (target) and 5.5 KB (target plus

insert) were subjected to the PCR amplification procedure (together) with the same outside primers, only the 1.5 KB piece would be amplified. Such a PCR strategy can be used to recover a true recombinant (without an insert) from a background of material with an insert. In an alternative mode of the embodiment, size selection of the PCR product can be achieved by agarose gel electrophoresis rather than by limiting the extension time for a PCR reaction. In this mode of the embodiment, the sequence inserted into or deleted from the target module can be of any size, with the constraint that the size difference between the PCR products of the unrecombined target module and the recombinant variant target module should be detectable.

The negatively selectable marker can be present in the target vector either alone or in conjunction with a positively selectable marker (preferably different from the positively selectable marker or markers present on the donor vector), for example, present as part of a selectable marker cassette. Such cassettes include, but are not limited to, Gal-Spec and Kan-Suc cassettes, as described in Section 6, below.

5.1.2.1.2 POSITIVELY SELECTABLE MARKERS

In another embodiment, recombinants may be selected by reconstruction of a flanking positively selectable marker. In this embodiment, the donor and target vectors are designed so that homologous recombination between regions of homology on the donor and target vectors result in the reconstruction of a sequence which encodes a selectable marker which was absent in both donor and target vectors. Thus, selection for the newly reconstructed selectable marker allows for selection of a recombination event that also results in the creating of a variant target gene sequence.

For example, in one embodiment of this method, outlined in FIG. 5, the target vector comprises a target recombination module comprising, in the following order from 5' to 3': a sequence **wxy**, a non-functional fragment of a sequence **wxyz**, which in turn encodes a functional gene product; a first target DNA sequence (**AB** in FIG. 5); and a second target DNA sequence (**c**) in FIG. 5). The donor recombination module comprises, in the following order from 5' to 3': a sequence **xyz**, which is a second non-functional fragment of the sequence **wxyz**; a first donor sequence (**ab** in FIG. 5); and a second donor sequence ("**c**") in FIG. 5). The first and second target DNA sequences are homologous to a sequences in the first and second donor DNA sequences, as described in Section 5.1.1.1, above. Recombination between donor and target sequences results in reconstruction of the sequence **wxyz**, which is able to encode a functional selectable marker. Thus, selection for the newly reconstructed selectable marker allows for selection of a recombination event, and selection against cells which lack recombinant vectors. Other variations of this method

include embodiments wherein the incomplete sequences are located immediately 3' to the second donor sequence and second target DNA sequences.

Further, any gene that conveys a readily identifiable or selectable phenotypic change, such as resistance to an antibiotic effective in *E. coli*, can be used as a selectable marker. Preferably, the selectable marker is an antibiotic resistance gene, such as the kanamycin resistance gene from TN903 (Friedrich and Soriano, 1991, Genes. Dev. 5:1513-1523), or genes that confer resistance to other aminoglycosides (including but not limited to dihydrostreptomycin, gentamycin, neomycin, paromycin and streptomycin), the β-lactamase gene from IS1, that confers resistance to penicillins (including but not limited to ampicillin, carbenicillin, methicillin, penicillin N, penicillin O and penicillin V). Other selectable genes sequences include, but are not limited to gene sequences encoding polypeptides which confer zeocin resistance (Hegedus *et al.* 1998, Gene 207:241-249). Other antibiotics that can be utilized are genes that confer resistance to amphenicols, such as chloramphenicol, for example, the coding sequence for chloramphenicol transacetylase (CAT) can be utilized (Eikmanns *et al.* 1991, Gene 102:93-98). As will be appreciated by one skilled in the art, other non-antibiotic methods to select for maintenance of the plasmid may also be used, such as, for example a variety of auxotrophic markers (see Sambrook *et al.*, 1989, *supra*; Ausubel *et al.*, *supra*), which can be selected by adding or subtracting a particular nutrient from the growth media.

5.1.2.2 ADDITIONAL TARGET VECTOR SEQUENCES

As described above for the donor vector, the target vector is compatible with all vectors present in the donor cell with respect to replication origin and the selectable marker and/or reporter genes, and is compatible with any other vectors residing in the target cell. Such requirements are described in Sections 5.1.1.3 and 5.1.1.4, above. As discussed in those sections, the chosen vector must be compatible with the donor vector plasmid described in Section 5.1, above. One of skill in the art will readily be aware of the compatibility requirements necessary for maintaining multiple plasmids in a single cell. Methods for propagation of two or more constructs in procaryotic cells are well known to those of skill in the art. For example, cells containing multiple replicons can routinely be selected for and maintained by utilizing vectors comprising appropriately compatible origins of replication and independent selection systems (see Miller *et al.*, 1992, *supra*; Sambrook *et al.*, 1989, *supra*).

Optionally, the target vector has additional features necessary for the screening or selection of the desired phenotypic characteristic of the recombined target gene, which is referred to herein as the "variant target gene", and which contains a "variant sequence module". In certain embodiments, for example, where screening or selection of

the desired phenotypic characteristic of the variant target gene is performed within the target cell itself, or where variant target gene products are purified from the target cell, signals for expression of the variant target gene, and/or a reporter gene construct may be included in the target vector. In an alternative embodiment, the variant target gene is transferred to a secondary host for screening or selection of the desired phenotype. In this embodiment, the target vector may contain sequences that allow transfer, maintanence or propagation of the vector in the secondary host cell (*e.g.*, mammalian tissue culture cells). For example, the target vector may include specialized origins of replication and expression systems, that allow expression of the variant genes in a secondary host. In one embodiment, for example, the target vector further comprises an SV40 origin of replication. FIG. 2 summaries these features.

In one embodiment, the target vector may contain sequences for regulating expression of the target DNA sequence, target gene, or variant target gene. With respect to regulatory controls which allow expression, either regulated or constitutive, at a range of different expression levels, a variety of such regulatory sequences are well known to those of skill in the art. The ability to generate a wide range of expression is advantageous for utilizing the methods of the invention, as described below. Such expression can be achieved in a constitutive as well as in a regulated, or inducible, fashion.

Inducible expression yielding a wide range of expression can be obtained by utilizing a variety of inducible regulatory sequences. In one embodiment, for example, the *lacI* gene and its gratuitous inducer IPTG can be utilized to yield inducible, high levels of expression of the target gene sequences, *e.g.*, a reassembled target gene sequence, when the sequences are transcribed via the *lacOP* regulatory sequences.

Preferably, the expression of a variant target gene is controlled by an inducible promoter. Inducible expression yielding a wide range of expression can be obtained by utilizing a variety of inducible regulatory sequences. In one embodiment, for example, the *lacI* gene and its gratuitous inducer IPTG can be utilized to yield inducible, high levels of expression of a target sequence, *e.g.*, a reassembled target gene, when sequences encoding such polypeptides are transcribed via the *lacOP* regulatory sequences. A variety of other inducible promoter systems are well known to those of skill in the art which can also be utilized. Levels of expression from reassembled target gene constructs can also be varied by using promoters of different strengths.

Other regulated expression systems that can be utilized include but are not limited to, the *araC* promoter which is inducible by arabinose (AraC), the TET system (Geissendorfer and Hillen, 1990, Appl. Microbiol. Biotechnol. 33:657-663), the p_L promoter of phage λ temperature and the inducible lambda repressor CI_{857} (Pirrotta, 1975, Nature 254: 114-117; Petrenko *et al.*, 1989, Gene 78:85-91), the *trp* promoter and *trp* repressor

system (Bennett *et al.*, 1976, Proc. Natl. Acad. Sci USA 73:2351-55; Wame *et al.*, 1986, Gene 46:103-112), the *lacUV5* promoter (Gilbert and Maxam, 1973, Proc. Natl. Acad. Sci. USA 70:1559-63), *lpp* (Nokamura *et al.*, *et al.*, 1982, J. Mol. Appl. Gen. 1:289-299), the T7 gene-10 promoter, phoA (alkaline phosphatase), *recA* (Horii *et al.* 1980), and the tac promoter, a trp-lac fusion promoter, which is inducible by tryptophan (Amann *et al.*, 1983, Gene 25:167-78), for example, are all commonly used strong promoters, resulting in an accumulated level of about 1 to 10% of total cellular protein for a protein whose level is controlled by each promoter. If a stronger promoter is desired, the tac promoter is approximately tenfold stronger than lacUV5, but will result in high baseline levels of expression, and should be used only when overexpression is required. If a weaker promoter is required in bacterial cells, other bacterial promoters are well known in the art, for example, maltose, galactose, or other desirable promoter (sequences of such promoters are available from Genbank (Burks *et al.* 1991, Nucl. Acids Res. 19:2227-2230).

In another embodiment, where it is desired to transfer the variant target gene into a secondary host for expression and screening assays, a target vector may also contain sequences for expression of the reassembled target gene in eukaryotic cells. Methods for the construction of such vector sequences may include *in vitro* recombinant DNA and synthetic techniques and *in vivo* recombinants (genetic recombination). Expression of nucleic acid sequence encoding a reassembled target protein or peptide fragment may be regulated by a second nucleic acid sequence so that the reassembled target protein or peptide is expressed in a host transformed with the recombinant DNA molecule. For example, expression of a reassembled target gene or gene product may be controlled by any promoter/enhancer element known in the art. Promoters which may be used to control reassembled target gene or gene product include, but are not limited to, the SV40 early promoter region (Benoist and Chambon, 1981, Nature 290:304-310), the promoter contained in the 3' long terminal repeat of Rous sarcoma virus (Yamamoto, *et al.*, 1980, Cell 22:787-797), the herpes thymidine kinase promoter (Wagner *et al.*, 1981, Proc. Natl. Acad. Sci. U.S.A. 78:1441-1445), the regulatory sequences of the metallothionein gene (Brinster *et al.*, 1982, Nature 296:39-42); plant expression vectors comprising the nopaline synthetase promoter region (Herrera-Estrella *et al.*, 1984, Nature 303:209-213) or the cauliflower mosaic virus 35S RNA promoter (Gardner *et al.*, 1981, Nucl. Acids Res. 9:2871), and the promoter of the photosynthetic enzyme ribulose biphosphate carboxylase (Herrera-Estrella *et al.*, 1984, Nature 310:115-120); promoter elements from yeast or other fungi such as the Gal 4 promoter, the ADC (alcohol dehydrogenase) promoter, PGK (phosphoglyceroyl kinase) promoter, alkaline phosphatase promoter, and the following animal transcriptional control regions, which exhibit tissue specificity and have been utilized in transgenic animals: elastase I gene control region which is active in pancreatic

acinar cells (Swift *et al.,* 1984, Cell 38:639-646; Ornitz *et al.* 1986, Cold Spring Harbor Symp. Quant. Biol. 50:399-409; MacDonald, 1987, Hepatology 7:425-515); insulin gene control region which is active in pancreatic beta cells (Hanahan, 1985, Nature 315:115-122), immunoglobulin gene control region which is active in lymphoid cells (Grosschedl *et al.*, 1984, Cell 38:647-658; Adames *et al.*, 1985, Nature 318:533-538; Alexander *et al.*, 1987, Mol. Cell. Biol. 7:1436-1444), mouse mammary tumor virus control region which is active in testicular, breast, lymphoid and mast cells (Leder *et al.*, 1986, Cell 45:485-495), albumin gene control region which is active in liver (Pinkert *et al.*, 1987, Genes and Devel. 1:268-276), alpha-fetoprotein gene control region which is active in liver (Krumlauf *et al.*, 1985, Mol. Cell. Biol. 5:1639-1648; Hammer *et al.*, 1987, Science 235:53-58; alpha 1-antitrypsin gene control region which is active in the liver (Kelsey *et al.*, 1987, Genes and Devel. 1:161-171), beta-globin gene control region which is active in myeloid cells (Mogram *et al.*, 1985, Nature 315:338-340; Kollias *et al.*, 1986, Cell 46:89-94; myelin basic protein gene control region which is active in oligodendrocyte cells in the brain (Readhead *et al.*, 1987, Cell 48:703-712); myosin light chain-2 gene control region which is active in skeletal muscle (Sani, 1985, Nature 314:283-286), and gonadotropic releasing hormone gene control region which is active in the hypothalamus (Mason *et al.*, 1986, Science 234:1372-1378).

In another embodiment, the target vector comprises sequences for transfer of the recombined vector carrying the variant recombination module to a secondary host organism for expression, screening, and/or selection assays. For example, so-called shuttle vectors have been designed to allow replication in a host bacterium, such as *E. coli*, and also allow transfer and replication in a variety of organisms, such as other bacteria (*e.g.*, Bruckner, 1992, Gene 122: 187-92); yeast (*e.g.*, Brunelli and Pall, 1993, Yeast 9: 1309-18); plants (*e.g.* Stanley, 1993, Curr. Opin. Genet. Dev. 3: 91-6); and mammalian systems (*e.g.* Karreman, 1998, Nucleic Acids Res. 26: 2508-10), where the subsequent selections can be performed. To act as a shuttle vector, the target vector should be able to replicate in the bacterial host to take advantage of both rapid generation times and, optionally, the simple genetic conjugation-based exchange systems. The target vector can readily be modified to include features of shuttle vectors, which are well know to those of skill in the art (see, *e.g.*, Pouwels, Cloning Vectors : a Laboratory Manual, Supplementary Update 1988, Elsevier; New York, NY 1988).

In yet another embodiment, the target vector comprises restriction endonuclease recognition sites to facilitate molecular manipulation of the variant target module, for example so that the variant target can be cloned into a different vector.

5.1.3 CELLS

Target host cells may be of any cell type which is capable of supporting homologous recombination. A cell capable of supporting homologous recombination contains a recombinase activity that catalyzes strand exchange between sequences with stretches of homology. In a preferred embodiment, the cells are bacterial cells and typically contain one or more bacterial recombinases. In embodiments where the donor vector is transferred to a bacterial target cell by conjugation of a bacterial donor cell with the bacterial target cell, donor and target host cells may be of any cell type which is capable of conjugative transfer of DNA. Such cells are well known to those of skill in the art. See, *e.g.*, Ely, B., 1985, Mol. Gen. Genet. 200:302-304. In embodiments where the donor vector is transferred to a target cell by conjugation, the target host cell is preferably naturally transformable to circumvent the need for preparing competent cells for transformation. In embodiments where the donor vector is transferred to a target cell by infection with a phage comprising the donor vector, the target cell must be capable of supporting transfer of donor sequences by the phage of choice. In a preferred embodiment, the phage comprising the donor vector is not capable of a full cycle of infection in the target cell, *e.g.*, cannot lyse a target cell into which a donor vector has been transferred.

Preferably, the target cell and, where utilized, the donor cell, are gram-negative bacterial cells, but gram-positive cells are also possible. More preferably, the host cell is an Enterobacterial cell. Members of the family *Enterobacteriaceae* include, but are not limited to, species of *Escherichia*, *Salmonella*, *Citrobacter*, *Klebsiellae*, and *Proteus*. Most preferably the host cell is an *Escherichia coli* cell. Naturally transformable bacteria for use with transformation-mediated transfer of the donor vector into the target cell include, for example, *Acinetobacter calcoaceticus*, *Haemophilus influenzae* and *Neisseria meningitidis* (Smith *et al.*, 1999, Res. Microbiol. 150(9-10):603-16). In embodiments where donor cells are utilized, the donor and target cells should comprise sequences or genetic backgrounds that allow independent selection for or against the presence of either the donor or the target cell. For example, the growth requirements and/or antibiotic resistance characteristics of the target and donor cells can be designed such that the presence of target cells can be selected for and/or the presence of donor cells can be selected against. Alternatively, methods for segregation of donor sequences can be utilized such as those described, below, in Section 5.2.3.

Target cells can also be derived from any organism, including, but not limited to, yeast, insect, or mammalian cells, provided they express, or can be engineered to express, a homologous recombinase activity capable of mediating recombination between two DNA molecules containing at least one region of sequence homology. The

recombinase is preferably a recombinase derived from *E. coli*. Such recombination-proficient cells may be made electrocompetent in advance and stored at -70°C.

5.1.4 DETERMINING SEQUENCE IDENTITY BETWEEN DONOR AND TARGET MODULES

As discussed above, the donor and the target sequences are homologous to each other. The extent of homology between the first donor sequence and the first target sequence, or between the second donor sequence and the second target sequence, is preferably at least 70% sequence identity. In other embodiments, the extent of sequence identity preferably at least 75%, 80%, 85%, 90% or 95% identity. In certain specific embodiments, the extent of sequence identity between donor and target sequences is at least 92%, 94%, 96%, 98% or 99%. A percentage of sequence identity between donor and target sequences that is 95% or greater, most preferably at least 98%, is desirable when the one-step selection method is utilized for selection of recombinant modules. Homologous sequences may be interrupted by one or more non-identical residues, for example for addition of novel sequences that can add function to a protein as described in Section 5.2.3, *supra*, provided they are still efficient substrates for homologous recombination.

To determine the percent identity of two nucleic acid sequences, the sequences are aligned for optimal comparison purposes (*e.g.*, gaps can be introduced in the sequence of the donor sequence for optimal alignment with the target nucleic acid sequence, particularly where one or both of the donor and target sequences are interrupted by extraneous sequences). The nucleotides at corresponding nucleotide positions are then compared. When a position in the donor sequence is occupied by the same nucleotide as the corresponding position in the target sequence, then the molecules are identical at that position. The percent identity between the two sequences is a function of the number of identical positions shared by the donor and target sequences (*i.e.*, % identity = # of identical overlapping positions/total # of positions x 100%). In one embodiment, the two sequences are the same length.

The determination of percent identity between two sequences can also be accomplished using a mathematical algorithm. A preferred, non-limiting example of a mathematical algorithm utilized for the comparison of two sequences is the algorithm of Karlin and Altschul (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*:2264-2268, modified as in Karlin and Altschul (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*:5873-5877. Such an algorithm is incorporated into the NBLAST and XBLAST programs of Altschul *et al.*, 1990, *J. Mol. Biol. 215*:403-0. BLAST nucleotide searches can be performed with the NBLAST nucleotide program parameters set, *e.g.*, for score=100, wordlength=12 to obtain nucleotide sequences homologous to a donor or target nucleic acid. To obtain gapped alignments for

comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.*, 1997, *Nucleic Acids Res. 25*:3389-3402. Alternatively, PSI-BLAST can be used to perform an iterated search which detects distant relationships between molecules (*Id.*). When utilizing BLAST, Gapped BLAST, and PSI-Blast programs, the default parameters of the respective programs (*e.g.*, of XBLAST and NBLAST) can be used (see, *e.g.*, http://www.ncbi.nlm.nih.gov). Another preferred, non-limiting example of a mathematical algorithm utilized for the comparison of sequences is the algorithm of Myers and Miller, (1988) *CABIOS 4*:11-17. Such an algorithm is incorporated in the ALIGN program (version 2.0) which is part of the GCG sequence alignment software package.

The percent identity between two sequences can be determined using techniques similar to those described above, with or without allowing gaps. In calculating percent identity, typically only exact matches are counted.

5.2 METHODS FOR DIRECTED GENE ASSEMBLY

The methods of the invention, as described in detail herein, can be used for a number of purposes, such as: 1) reassembling genes from sequence-related members of gene families; 2) site-directed mutagenesis; 3) inserting or substituting sequences in a target gene to construct recombined vectors; and 4) combinations of these processes. Variant sequences resulting from any of these processes can, for example, be archived and/or tested for optimization of a desired phenotype. These methods are described in detail herein.

In general, the DGA method comprises the steps of: transferring a donor vector, optionally contained within a donor cell, as described in Section 5.1.1, above, into a target cell having a target vector containing a target gene or gene sequence of interest, as described in Section 5.1.2, above, allowing homologous recombination to occur between the donor vector and the target vector, and selecting for a target cell containing a variant of the target gene of interest. Conditions that allow homologous recombination to occur merely refer to standard growth or maintenance conditions for the particular cells being used in the particular instance. As also discussed above, the target gene or gene sequence of interest can, in an alternative embodiment, be integrated into the genome of the target cell.

Prior to the step of transferring the donor vector into the target cell, the donor sequences may be subjected to any of a variety of mutagenesis procedures in order to produce a pool of diverse donor sequences. A schematic of this strategy is shown in FIG. 6. Mutagenesis procedures are well known in the art. In one embodiment, donor vectors may be mutagenized either *in vitro*, prior to introduction into donor cells by *in vitro* mutagenesis protocols (*e.g.*, Edward, 1996, Methods Mol. Biol. 57: 97-107). In another embodiment, donor vector may be mutagenized in *in vivo*, for example using *E. coli* mutator strains (see

e.g., Horst et al., 1999, Trends Microbiol., 7:29-36; Miller and Michaels, 1996, Gene 179:129-32; Miller, 1998, 409:99-106). Some non-limiting examples of such mutator strains are mut D, mut S, mut Y, and mut M.

As described in Sections 5.2.1 and 5.2.2 below, selection for a target cell containing a variant gene can be accomplished be a one-step method or, preferably when the percentage sequence identity between the donor and target is less than 95%, a two step method. The one-step method selects for the product of the homologous recombination, *i.e.* a variant target gene. This selection can be direct or indirect, the former entailing selection for recombined sequences and the latter entailing selection against unrecombined target sequences. The two-step selection method entails, prior to selection for the variant target gene, the additional step of selecting for the intermediate of homologous recombination, a structure known as a co-integrant. Following selection of variant target molecules, the donor sequences can be segregated as described in Section 5.2.3, *infra*.

It is noted that multiple selections can be performed at any of the selection steps. For example, appropriate target and donor vectors can be designed such that a multiple selection for loss of a negatively selectable marker and a molecular selection (*e.g.*, using amplification to select for a particular size of sequence) can be performed. Multiple selections can make possible the identification and isolation of particularly rare events such as, for example, identification and isolation of a somatic mutation in a population of wild type allele copies. A representative, non-limiting example of multiple selection is demonstrated in Section 6.5, below.

5.2.1 ONE-STEP SELECTION OF VARIANT TARGET MOLECULE

Homologous recombination between the donor recombination module of the donor vector and the recombination module of the target vector gives rise to a variant target molecule. The one-step selection method of the variant target described hereinbelow is preferably used where the target and donor sequences being recombined share at least 95% sequence identity. Recombinant products may be selected in a number of ways, depending on the choice of selectable markers in the target vector, as described above in Sections 5.1.2.1.1 and 5.1.2.1.2. As described therein, recombinant variant modules maybe selected by placing sequences that are detrimental to cell growth under a controlled set of conditions, so-called conditional lethal sequences, within a region targeted (see Section 5.1.2.1.1), or by the elimination of a polar sequence (see Section 5.1.2.1.1).

Because the target vector has a negative selection marker between the first target sequence and the second target sequence, selection of a variant target molecule can simply entail selection against the negative selection marker, which is lost as a result of the homologous recombination process. Thus, in one embodiment, selection for recombinants

is by a negative selection method, as described above in Section 5.1.2.1.2, above. This method comprises the steps of: (a) transferring a donor vector into a target cell, *e.g.*, a bacterial cell, which is capable of homologous recombination, wherein (i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first donor DNA sequence and a second donor DNA sequence, and (ii) said target cell comprises a target vector comprising a target recombination module comprising, in the following order from 5' to 3': a first target DNA sequence; a negatively selectable marker; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence; and (b) selecting for a population of target cells which do not contain the negatively selectable marker, so that a population of a variant sequence modules in cells is generated. The cells undergoing DGA are subjected to conditions that allow homologous recombination to take place. Conditions that allow homologous recombination to occur merely refer to standard growth or maintenance conditions for the particular cells being used in the particular instance. Such conditions are well known to the skilled artisan.

Generally, selecting for target cells that do not contain the negatively selectable marker is accomplished by subjecting the cells to conditions that do not allow growth of donor cells or of target cells that still contain the negatively selectable marker (*i.e.*, have not undergone recombination with the donor vector resulting in loss of the negatively selectable marker). To ensure loss of donor cells, for example, a selectable marker (*e.g.*, a tetracycline resistance-encoding element) can be included in the chromosomal background of the target cell, but be absent from the donor cell. Imposing appropriate selective pressure (*e.g.*, inclusion of tetracycline) results in selected loss of donor cells. In a variation of this method, the target recombination module is present in the target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

In an alternative method, a positive selection method, as described above in Section 5.1.2.1.2, above, is used to select for recombinants. In this case, a first non-functional fragment of a positively selectable marker flanks the donor recombination module, and a second non-functional fragment of the marker flanks the target recombination module. Appropriate recombination between the marker fragments and between the donor and target recombination modules results in reconstruction of a function marker. Thus, selection for the presence of a functional positively selectable marker selects for a recombinant target gene of interest. This method comprises the steps of: a) transferring a donor vector into a target cell, *e.g.*, a bacterial cell, which is capable of

homologous recombination, wherein i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first non-functional fragment of a positively selectable-marker; a first donor DNA sequence; and a second donor DNA sequence; ii) said target cell comprises a target vector comprising a target recombination module comprising, in the following order from 5' to 3': a second non-functional fragment of the positively selectable-marker; a first target DNA sequence; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence, and recombination between said first non-functional fragment of the positively selectable-marker and said the second non-functional fragment of the positively selectable-marker results in a functional positively selectable marker; and (b) selecting for a population of target cells which contain the positively selectable marker, so that a population of a variant sequence modules in the cells is generated. In a variation of this method, the target recombination module is present in the target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

5.2.2 TWO-STEP SELECTION OF VARIANT TARGET MOLECULE

In another embodiment, a two-step procedure is used to select for the product of homologous recombination, which entails selection of an intermediate state in the process followed by selection of the product of homologous recombination. In such an embodiment, the intermediate state is one in which the target cell contains both the donor vector and the target vector. Without wishing to be bound by any theory or mechanism, it is believed that this intermediate state more particularly involves an intermediate of the homologous recombination process referred to as a co-integrant. In the latter embodiment, a fourth element is required, namely a positively selectable sequence in the donor DNA to allow for selection of the intermediate state. This sequence can be present at any position of the donor vector that does not interfere with standard vector functions (*e.g.*, vector replication).

The invention encompasses, first, a method for generating a population of variant sequence modules in cells, *e.g.*, bacterial cells, said method comprising: (a) transferring a donor vector into a target cell which is capable of homologous recombination, wherein (i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first donor DNA sequence and a second donor DNA sequence, and additionally comprises a positively selectable marker; and (ii) said target cell comprises a target vector comprising a target recombination module comprising, in the

following order from 5' to 3': a first target DNA sequence; a negatively selectable marker; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence; (b) selecting for target cells that contain the positively selectable marker; and (c) selecting for a population of target cells which do not contain the negatively selectable marker, so that a population of variant sequence modules in cells, in particular, the target cells, is generated. Generally, selecting for target cells that do not contain the negatively selectable marker is accomplished by subjecting the cells to conditions that do not allow growth of donor cells or of target cells that still contain the negatively selectable marker (*i.e.*, have not undergone recombination with the donor vector resulting in loss of the negatively selectable marker). To ensure loss of donor cells, for example, a selectable marker (*e.g.*, a tetracycline resistance-encoding element) can be included in the chromosomal background of the target cell, but be absent from the donor cell. Imposing appropriate selective pressure (*e.g.*, inclusion of tetracycline) results in selected loss of donor cells. In a variation of this method, the target recombination module is present in the target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

In another embodiment, the invention provides a method for generating a population of a variant sequence modules in cells, *e.g.*, bacterial cells, said method comprising:(a) transferring a donor vector into a target bacterial cell which is capable of homologous recombination, wherein (i) said donor vector comprises a donor recombination module comprising, in the following order from 5' to 3': a first non-functional fragment of a first positively selectable marker; a first donor DNA sequence; and a second donor DNA sequence, and additionally comprises a second positively selectable marker; (ii) said target cell comprises a target vector comprising a target recombination module comprising, in the following order from 5' to 3': a second non-functional fragment of the positively selectable marker; a first target DNA sequence; and a second target DNA sequence, wherein said first donor DNA sequence is homologous to said first target DNA sequence, and said second donor DNA sequence is homologous to said second target DNA sequence, and recombination between said first non-functional fragment of the selectable marker and said second non-functional fragment of the selectable marker results in a functional selectable marker; (b) selecting for target cells that contain the second positively selectable marker; and (c) selecting for a population of target cells which contain the first functional positively selectable marker, so that a population of a variant sequence modules in the cells is generated. In a variation of this method, the target recombination module is present in the

target cell integrated into the target cell genome. Preferably, the target recombination module is integrated in a manner that readily allows excision or isolation of the module out genome, *i.e.*, via flanking unique restriction sites or by specific amplification of the module.

With respect to co-integrants, without wishing to be bound by any theory or mechanism, co-integrant formation is driven by homologous recombination in regions of shared homology. Co-integrants are intermediates of homologous recombination that can be selected for by subjecting target cells into which a donor vector has been transferred to conditions that select for a marker present on a target vector. Co-integrants are unstable in the absence of selective pressure. Co-integrant structures can resolve in one of two different ways; the reverse reaction yields the original donor and target, and a forward reaction produces a variant target. Without wishing to be limited by any particular theory or mechanism, in the methods described herein, it is believed that placement of the negatively selectable marker in the target and subsequent selection against said marker drives the recombination event in the forward direction. Recombination between regions of homology either side of the site of the negative selection insert will lead to a recombination event that directs the assembly of the gene with the desired new segment of DNA. FIG. 14 shows how the process of co-integrant formation followed by DGA-selected resolution breaks the process illustrated in FIG. 3 into sequential and separable steps.

In a preferred version of this embodiment, the donor vector is a suicide vector (see Section 5.2.3, below) that replicates only in the donor cell, not the target cell. Use of a suicide vector, coupled with selection for the second positively selectable marker favor co-integrant formation.

The selection for and maintenance of co-integrants can be useful in generating diversity, as a single co-integrant can give rise to a family of recombinant molecules. Specifically, selection for co-integrant formation selects for a first recombination event, and co-integrant resolution can be accomplished via recombination at a number of positions, thereby creating a family of sequence variants. A representative example of this is presented, below, in Section 6.5.1.

Thus, in one embodiment of the present invention, selection of a variant target molecule comprises two steps. In the first step, selection for the co-integrant is achieved by selecting for a positively-selectable marker on the donor vector. In the second step, selection for unrecombined target vectors is achieved as described in Section 5.2.1 above, for example, by selecting against a negatively-selectable marker in the target module.

5.2.3 SEGREGATION OF DONOR SEQUENCES

In certain embodiments of the invention, selection for the segregation of donor sequences, that is, loss or removal of unrecombined donor sequences and non-recombination module sequences, after selecting for cells containing recombinant variant modules is desired. In one embodiment, both replication functions and transfer functions are provided from genes provided *in trans* to the donor vector to prevent replication and transfer of the donor vector following in the target cell (Metcalf *et al.*, 1994, Gene 138:1-7). Where a reciprocal homologous recombination event replaces a conditionally lethal, negatively-selectable marker, recombination may result in exchange of the conditionally lethal marker to a second replicon. If the second replicon has a conditional origin of replication, then loss of the counter-selected marker can be facilitated by conditions that are incompatible with replication of the second replicon (Penfold and Pemberton, 1992, Gene 118:145-6). This strategy is outlined in FIG. 10. In a preferred embodiment of the present invention the donor vector replicates in the donor cell but fails to replicate in the target cell. The use of such a suicide vector facilitates selection for recombinants when selecting for a negatively selectable marker, because the donor vector is lost from the target cell following recombination.

5.2.4 PHENOTYPE OPTIMIZATION

Once sequence variants are generated, the variant sequences or genes can be screened and optimized for a desired phenotype of interest. The selection process drives the optimization of the sequence or gene during iterative rounds of the process. The selection method chosen will be depend on the nature of the target sequence and the desired property to be optimized, and will be apparent to the skilled artisan in the particular area of interest. The sequences can be subjected to any selective pressure appropriate to optimize the particular phenotype of interest. Selection can occur in the target cells containing the variant sequences, or can be performed in a secondary cell type, either in culture or *in vivo*. Representative, non-limiting, examples of the types of phenotype optimization the DGA methods of the present invention can be used in conjunction with are presented hereinbelow.

In one embodiment, for example, the variant target sequence may encode a transcription factor for which the property of increased ability to activate a particular target gene is desired. In this case, the selection system could comprise a reporter gene, operatively linked to a transcription factor binding site, such that binding of the transcription factor results in expression of the reporter gene. The assay can comprise identifying a variant transcription factor that results in increased activation of the reporter gene relative to the target gene, *i.e.* the wild-type transcription factor. Such an assay may

be accomplished either in the target cell itself, or the recombinant variant gene may be transferred to a secondary host cell for expression and selection.

Alternatively, the variant target sequence can encode an enzyme whose activity is to be optimized (*e.g.*, substrate can be modified and/or activity can be increased or attenuated). For example, in the case of industrial enzymes, the phenotype of an enzyme of interest can be subjected to appropriate selective pressure to optimize such phenotypic properties as the substrate specificity, temperature resistance, salt tolerance, pH range, or solvent tolerance or otherwise extend the environmental parameters under which enzymes that have industrial applications, including but not limited to, proteases, esterases, oxidases, dehydrogenases, catalases, lactases, or other such enzymes function.

In the agricultural area, for example, variant target sequences can be subjected to appropriate selective pressures to optimize properties of food storage proteins to improve quality traits of a crop. For example, the DGA methods of the present invention can be utilized to alter genes encoding pathogen resistance determinants to extend the range of pathogen resistance, or to modify sequences involved in *e.g.*, salt, drought, and temperature tolerance to modify (*e.g.*, enhance) growth characteristics of a plant of interest.

In the medical area the DGA methods of the invention can, for exampe, be used to optimize antibody characteristics (*e.g.*, enhance, modifiy antigen specificity and/or improve binding), to produce a large pool of antibody diversity *in vitro*, and/orto humanize antibodies, *e.g.*, rodent antibodies. Further, proteins or polypeptides exhibiting therapeutic efficacy or potential can be optimized via the DGA methods of the present invention. For example, enzymes with therapeutic applications can have their reaction parameters made more amenable to the particular therapeutic situation. Further, proteins, *e.g.*, growth factors, can be optimized to beneficially alter efficacy, production or range of biological activities. Still further, the DGA methods of the present invention can be used to reduce the immunogenicity of protein therapeutics, or to enhance the antigenicity of immunizing antigens.

5.2.5 DIRECTED GENE ASSEMBLY VARIATIONS

The DGA methods described in Section 5.2 sets forth the basic elements of DGA. Presented in this section are a few of the variations or modifications of the basic DGA method, *e.g.*, methods for production of complex populations of variants or variants having additional sequences that do not solely correspond to sequences homologous to sequences originally present in the target modules, that can also be routinely practiced.

For example, in one embodiment, a target recombination module comprises more than one negatively selectable marker, in order to direct recombination into more than one region of the target vector. An example of a target vector comprising two negatively

selectable markers, galE and sucB, in a single target vector is illustrated in FIG. 7. FIG. 7 shows a target recombination module with two negatively selectable markers inserted into the coding sequence of the target gene. Such sequences can be constructed by any method described herein for construction of target recombination modules containing a single negatively selectable marker, or by standard techniques well known in the art. For example, methods can be employed that comprise cloning into available restriction sites or transpositional insertion using insertion elements containing a negatively selectable marker.

In a non-limiting example of such an embodiment, first, a population of bacteria containing a target vector is mixed with a population of bacteria containing a library of donor vectors comprising gene or sequence fragments from a variety of genes related to the target gene (that is, represent homologs of, or at a minimum, exhibit sufficient sequence homology to allow homologous recombination with target gene sequences). In FIG. 7, two each of two family members is used for illustration purposes. Selection against the donor cells using the first negatively selectable marker on the target vector (in the example, gal) selects for and thereby produces recombinant molecules. Specifically, each donor vector can recombine with the target vector, resulting in replacement of the sequence flanking the first negatively selectable marker sequence by donor DNA. Different variant sequences are produced in each case, provided there is variation, *e.g.*, allelic variation, between different exchange points. This principle is illustrated in FIG. 7, showing two possible products with each donor vector. The product of the first exchange event still contains the second negatively selectable marker and, therefore, the target gene product itself still cannot be expressed. Nonetheless, these intermediate variant sequences can be produced and selected for because selection was exerted for a recombination event, independent of the nature of the target gene product, illustrating one of the advantages of the DGA methods of the present invention.

The variant target sequences generated via the first exchange can then become substrates for a second set of homologous recombination exchanges. (It is noted that, alternatively, such sequences can be archived for future use in another DGA application.) The second round of recombination produces and selects for taget recombination modules that have undergone recombination to lose the second negatively selectable marker. If desired, the resulting variant target sequences can then be expressed to assess the properties of the variant target gene product produced therefrom. This procedure is illustrated with one of the products of the first exchanges illustrated in FIG. 7. As FIG. 7 demonstrates, each individual member of the first exchange has the potential to produce an array of variant sequences. The procedure employed, therefore, provides for the combinatorial amplification of variant sequences.

FIG. 7 illustrates the process with a single target and two donors. Larger libraries of donors and/or targets can be used to produce vastly larger ensembles of product molecules. It is also possible to more carefully control the process by restricting the size of the donor DNA sequences, thus restricting the extent of the regions participating in the exchanges. The final products in such a strategy can, for example, be achieved employing a sequential procedure wherein a single negatively selectable marker is employed for the first product series and an intervening step is used to introduce a second collection of negatively selectable markers prior to the second round of targeting as described below, and illustrated in FIG. 8.

In another embodiment, the DGA methods of the present invention can be used to insert a heterologous sequence into a target gene or sequence, or replace a target gene or sequence with a heterologous gene sequence. The recombination events replacing the negatively selected insert require homology flanking both sides of the insert. Just as flanking homology can delete intervening non-homologous material, flanking homology can be used to introduce non-homologous sequences as inserts into a sequence, or as substitutions in a deletion-insertion process replacing existing segments of DNA. The fundamentals of such a procedure are illustrated in FIG. 9. Insertion of sequences is useful, for example, for introducing novel sequences that can add function to a protein, *e.g.* a second activity in a sequential enzyme pathway or specific cellular localization functions. For example, sequences encoding additional protein domains can be introduced into the coding region of a target gene sequence of interest. Further, additional selectable markers can be introduced into a target gene or sequence of interest via such an embodiment, thereby creating or modifying a target vector.

In addition to insertions of new sequences, a directed homologous recombination event can be used to replace segments of the target recombination module with sequences from the donor recombination module. The substitution process can, for example, execute the combinatorial replacement of sequences that are structural homologs of segments, *e.g.*, segments in a gene family, that may fail to have the sequence homology required for the direct homologous replacement in a re-assortment process. For example, such an embodiment can result in "domain swapping," that is, sequences encoding a particular domain can be substituted for sequences encoding a different, either related or unrelated, domain. Such structurally related stretches, with low homology will also in many instances fail to provide adequate substrates for PCR re-assortment strategies. Insertional substitution can substantially extend the scope of sequences that can be directed to participate in a combinatorial re-assortment process.

Still further, the DGA methods of the present invention can also be used to generate new target vectors by, for example, moving negatively selectable markers from

donor vectors to target vectors. To accomplish this, selectable markers are placed in the target recombination module. The resulting vector can then be used as a donor vector in a DGA procedure to new target vectors. A representative example of this is demonstrated in Section 6.5.3, below.

This process of "variant"-variant target recombination module production can be iterated any number of times by performing genetic crosses without *in vitro* manipulations. The process both uses and can produce reagents that may be archived. The method is illustrated in FIG. 8.

In another embodiment, the DGA methods of the present invention can be used to isolate specific sequences of interest from a library of sequences. For example, a library of donor vectors can be presented to a collection of target cells containing a target vector with a target recombination module. Selection can be designed such that the only cells allowed to grow are those target cells that have undergone selection with a donor DNA sequence. Because such a recombination event requires a minimum amount of homology, such a scheme serves to identify sequences within the library that contain homologous sequence. Thus, DGA allows, for example, evolutionary re-assortment from libraries of sequences without the need for prior identification and isolation of homologous candidate sequences. Further, the donor DNA sequences need not have extensive homology with the target DNA sequences, as long as sufficient homology exists to support homologous recombination. Limited homologies across gene segments are sufficient, especially when cells, *e.g.*, mutL cells, that lack mismatch repair function, are utilized. Such an embodiment can be used, for example, to capture homologous domains from otherwise dissimilar proteins. This strategy is illustrated in FIG. 11.

Multiplexing embodiments of the DGA methods of the present invention can readily be practiced. For example, DGA can be used to produce sequences that encode new proteins, by, *e.g.*, replacing particular structural motifs in a target protein with new sequence, using a DGA re-assortment or insertional substitution strategies. The context of the structural motif is likely to be important, however, and adjustments may be required to create a functional polypeptide. However, using conventional protocols, the suitability of the structural motif in the new context of the novel protein can be evaluated in one context at a time. By combining mutagenesis of the donor or target vector with a re-assortment or insertional substitution procedure, multiple novel proteins comprising an array of variants in a variety of contexts can routinely be evaluated.

The multi-component nature of the DGA process lends itself to combinatorial strategies. These combinatorial strategies can take place over an extended period of time and components of the process, because they are actual living replicating entities – cells, *e.g.*, bacteria, containing the donor and target vectors- may be archived (see

below) and amplified as desired in subsequent iterations of an experimental series. It is also possible that a target gene produce may have a variety of potential evolutionary endpoints, in which case, entire sets of vectors, *e.g.*, target vectors, can be reused in subsequent series of phenotype optimization experiments with different goals and results based on the application of different selective pressures and different subsequent direction of further sequence variation via DGA.

Conjugational gene transfer, a preferred procedure for transfer donor DNA into a target cell, is amenable to automation. Using liquid handling automation individual members of a donor library can be arrayed. Again employing liquid handling automation, an arrayed collection of donors may be individually mixed with a target. The behavior of the products resulting from the DGA exchanges can then be determined for an arrayed collection of products with reference maintained to the original donors that produced individual targets.

DGA can be used to query a domain or structural motif to see if it can substitute for an existing sequence in a target protein. To query a candidate sequence, a negatively selectable marker is be placed into a segment encoding the portion of the test protein with the domain (or structural motif) in question. DGA is then be used to drive the recombination process. If the queried candidate sequence has sufficient homology to drive the process it can be recombined directly from a donor vector. If the queried candidate sequence is of distant homology or a non-homologous structural homologue (candidate), it can be embedded into homologous sequences flanking the selectable marker as described above. In either instance counter-selection against the targeted selectable marker can be used to drive the recombination process directing the gene assembly. DGA drives the production of the gene product and the product can be tested and compared with the parental gene (summarized in FIG. 8). Relative activities (defined by the specifics of the test protein) define the relative ability of the candidate sequence to substitute for the domain (or structural motif) in the test protein. It is also possible to combine the above procedure with mutagenesis to assess the "sequence space" neighboring the precise input combination. In this way information about the full potential of the queried motif in the new context can be derived.

5.3 LIBRARIES

The invention further provides libraries suitable for the practice of directed gene assembly. Such libraries can be donor or vector libraries and can comprise a plurality of any of the donor or target vectors of the invention, including vectors comprising variant target sequences that have been produced via DGA. Such libraries can also comprise variant target gene or target gene sequences produced via DGA that no longer contain

intervening selectable markers and encode variant target gene products, including optimized variant target gene products. Such libraries can also comprise cells containing co-integrant configurations that can, at a desired point, be resolved. Libraries can also comprise a plurality of archived sequences or modules, (see Section 5.4, below) optionally present within cells.

In one embodiment, the vectors of the library are present within cells (*e.g.*, donor cells for donor vectors and target cells for target vectors), *e.g.*, bacterial cells.

In another embodiment, the library vector, preferably the donor library vectors, is a suicide vector. In yet another embodiment, the vectors of the donor libraries contain defective conjugative transfer sequences. In another embodiment, such vectors are present in donor cells that complement the conjugative transfer sequence defect.

In still another embodiment, the members of the library are arrayed. In another embodiment, the members are arrayed in a 96 (*e.g.*, 8 x 12), 384 *(e.g.*, 16 x 24), or a 1536 (*e.g.*, 32 x 48) matrix or plate, *e.g.*, microtiter plate. In yet another embodiment, the donor library is present in a multiplicity of cells, *e.g.*, bacterial cells, each cell containing a member of the library the members of which are arrayed. In another embodiment, such cells are arrayed in a 96 (*e.g.*, 8 x 12), 384 *(e.g.*, 16 x 24), or a 1536 (*e.g.*, 32 x 48) matrix or plate, *e.g.*, microtiter plate.

Donor vectors generally have a greater potential for subsequent reuse than target vectors. Firstly, the requirements of the donor vector are less constraining as there are no requirements for second origins (as with shuttle vectors often utilized as target vectors) or expression-related sequences. Donor libraries, therefore, can be universal as they can be be compatible with many target libraries. In such universal donor library embodiments, it is generally preferable to use smaller donor DNA sequences as such sequences are more likely to be useful and usable in multiple proteins.

For example, proteins are comprised of a finite variety of structural motifs (Thornton *et al.*, 1999, J. Mol. Biol. 293: 333-42). Sequences encoding motif-sized pieces in a donor library, for example, are likely to have uses in a large variety of proteins. It is possible to directly pursue the acquisition of a collection of protein motifs in a specialized donor library as described, above, in Section 5.3.

Further, as discussed above, in Section 5.3, homology-based isolation of gene sequences from libraries is a powerful application of DGA technology. Using this application a collection of gene sequences can be created in a donor vector producing a plurality of potential donor sequences. Such pluralities can have many applications across a variety of targets and, hence, represent valuable libraries and archives (see below). For example DNA from an extremeophile such as a thermophilic organism can be used to construct a library that is screened for sequences able to replace segments in enzyme X,

based on homology, with the goal of arriving at a more thermal resistant enzyme X. In addition, once made, such a library can prove useful for many subsequent experimental series with other enzymes.

In one donor library embodiment, therefore, the donor vectors of the library comprise related donor DNA sequences. For example, in such a library, the donor DNA is derived from: different homologs of the same gene or gene portion from different species; different members, or portions thereof, of a particular gene family exhibiting amino acid similarity; or different DNA sequences encoding polypeptide domains exhibiting amino acid similarity.

When products with desired properties are identified, the donors that were used in those specific crosses can be isolated and set aside to produce a specialized extracted library (FIG. 12). An extracted library is a library containing modules or sequences of similar or related function. The sequences of such an extracted library are likely to provide similar function or functions to proteins. Members of such extracted libraries can, for exmple, be accumulated during the course of experiments with specific gene product goals. Extracted libraries can also be produced as part of studies designed to isolate protein building blocks (structural motif or domains) for use in phenotype optimization and directed evolution experiments employing DGA strategies. Extracted libraries (regardless of the means used to assemble them), therefore, provide preformatted donor reagents that have described uses in specific contexts and, as such, can also represent archived modules (see the next section).

5.4 ARCHIVES

Discussed above and in the examples provided herein are methods and compositions relating to target vectors, donor vectors, and DGA methods. The present invention is also directed to archived sequences of any sequence or module produced via such methods. An archived module, as used herein, refers to a donor DNA sequence or target DNA sequence, whether or not the target sequence has undergone DGA or phenotype optimization, where the sequence comprising the archived module is known or has been demonstrated to encode a protein segment or domain that provides a particular function (*e.g.*, ligand binding, enzymatic activity, structural activity), and has been stored and catalogued (archived), *e.g.*, for future use, such as future use in similar or different DGA situations. The size and numbers of archived modules, and the information associated with the archives limited only by the number of experiments performed.

The bi-molecular nature of the methods described herein allows reagents to be used repeatedly, *e.g.*, as part of a sequential combinatorial process. It also allows the reagents, once created, to be archived. One of the principle advantages of the DGA

approach is, in fact, the ability to recycle reagents in subsequent iterations of an experiment. This can be extended beyond a simple set of experiments across many experiments creating an archive of reusable reagents. Many different types of archives are possible ranging from target and donor libraries simply frozen for potential future use, to extracted collections with proven function or use, and extending to archives of structural motifs and domains deliberately isolated as building blocks for rational protein design.

With time and multiple iteration of the process, both within a specific set of experiments and across many different experiments, information about the archived modules is built. Preferably, therefore, archived modules have a history relating to their behavior in previous DGA procedures. That is, in addition to the module itself, there is a store of information relating to the sequence and function history of the archived module. This history grows over time and allows subsequent DGA iterations or projects to be directed by the information accumulated. That is, it is such a history in a related series of experiments that can form the data, or part of the data, analyzed to direct iterative rounds of variant sequence production and phenotype optimization (directed evolution).

For example, in a particular round of DGA, the modules exchanged represent homologous segments of proteins, or at least contain flanking areas of homology. New combinations represent new re-assortments of structural components. Information about how a particular sequence behaves in a given context, or which sequences are functional or optimal in specific context(s), accumulates, and over time provides a database with information about the structural domains and motifs of the proteins involved that describe their use or activity, therefore, suggesting futures uses for the sequences in subsequent phenotype optimization and directed evolution. It is noted that this capacity to produce such archived module collections with associated data further distinguishes the methods of the present invention from random complex permutation sampling approaches to directed evolution.

Archived modules can routinely be frozen and cataloged. In a preferred embodiment, the archived module is present as part of a vector (generally a donor or target vector, with a donor vector being preferred). In another preferred embodiment, the archived module is present within a cell.

Where the donor vector is contained in a host bacterium for conjugation-mediated transfer, dramatic miniaturization can be employed as a single nanoliter of material contains 10^3 organisms. The growth rate of bacteria (1 generation every 15-20 minutes) allows aliquots to be amplified by a factor of 10^6 in six hours, and can permit 75 generations of "evolution" to be achieved each day. Simple liquid handling robotic systems can be employed to distribute and mix bacterial populations permitting the plasmid-based donor/target approach to take full advantage of the developments in high throughput

screening technologies that were achieved in the 1990s (See, *e.g.*, Cox *et al.*, 2000, Prog. Med. Chem. 37: 83-133).

5.5 DATABASES

The DGA approaches of the invention generate data relating to, *e.g.*, the behavior of structural motifs and protein domains as, for example, discussed above for archived modules. Such information represents a database of information. As such, the present invention still further provides a computer readable medium having a database recorded thereon in computer readable form, wherein said database comprises one or more module profiles and wherein each module profile describes a phenotype in a DGA assay, and wherein each module profile is associated with a particular vector in a particular target cell.

For example, if the donor input materials are arrayed, the results obtained about the arrayed individual products can be used (see above) to produce extracted libraries. The assembly of extracted libraries with modules of predefined uses will allow the "directed evolution" process to be directed, not only by the results of iterative screening and selections, but also by accumulated knowledge about extracted libraries and our growing understanding of protein structure. The DGA strategy of the present invention naturally lends itself to an eventual integration of directed evolution technologies with the theoretical developments in the field of rational protein design (Regan, 1999, Curr. Opin. Struct. Biol. 9: 494-499).

6. EXAMPLES

The following examples demonstrate construction of a donor vector series (Section 6.1) and a recipient donor series (Section 6.2) into which bacterial subtilisin genes were cloned, subjecting the bacterial subtilisin genes to DGA and two-step selection of variants: first, selection of co-integrant (Section 6.4), and selection of variant modules (Section 6.5). Section 6.6 demonstrates that the foregoing procedures resulted in the generation of a collection of functional variants of subtilisin molecules.

6.1 DONOR VECTOR

6.1.1 THE CREATION OF THE pGPG PLASMID SERIES

A universal pre-donor plasmid, pGPG, was designed for use with the DGA-related subject matter described herein. Briefly, the pGPG plasmid was designed to have: 1) a minimum amount of vector sequence homologous to other standard vectors; 2) a positively selectable marker; and 3) a multiple cloning site into which donor sequences of

interest can easily be introduced. As noted below (Section 6.2.4) such a vector can also be utilized in the construction of target vectors.

The pGPG plasmid is a derivative of the R6K plasmid. The plasmid R6K can be transferred between strains by conjugation (Macrina *et al.*, 1974, J. Bacteriol. 120(3):1387-1400). A significant number of derivatives of R6K have been created, among which are plasmids defective for conjugation (Nunez *et al.*, 1997, Mol. Microbiol. 24:1157-68), replication (Kolter, 1981, Plasmid 5(1):2-9), or for both conjugation and replication (Metcalf *et al.*, 1994, Gene 138:1-7). The plasmids can be rescued by providing the conjugation and/or replication functions in trans. An R6K derivative where replication and conjugation functions are provided in trans is desirable as a donor vector. Once such a derivative is transferred to a target strain which lacks the replication and conjugation functions, the vector DNA exists transiently pending dilution following bacterial growth. The vector DNA is available for recombination, but (in the absence of recombination) will rapidly be lost and will not replicate or participate in subsequent conjugational events. One such plasmid, pGP704 (http://salmonella.org.vectors/pgp704/), was used as starting point for the creation of the pGPG series of vectors suitable for DGA.

To eliminate sequences from the donor common to most commonly utilized vectors and, at the same time provide a useful selective marker, the plasmid pGP704 was partially digested with Bam HI to produce a 2216 base pair fragment which was ligated with a 865 base pair Bam HI fragment from the plasmid p34SGM (Dennis and Zylstra, 1988, J. Applied Environmental Microbiology 64(7):2710-2715) containing the *aacC1* gene and its promoter encoding the function conferring resistance to gentamycin resistance. The resultant ligation mixture was transformed into the π replication proficient host OTG28 (for all strains referred to herein, see Section 6.3, below) and plated on Luria agar selecting gentamycin (10 μg/ml) to isolate the plasmid pGPG6 (FIG. 15).

Further modifications to the pGPG6 were made to produce cloning vectors with unique multiple cloning sites ("MCS"; MCS1, pGPG7 and MCS2, pGPG8). pGPG6 was first cut with SmaI and Sac I, terminal nucleotides were removed (Sac I site) and the resultant molecule was circularized with ligase to produce pGPGSS (FIG. 15). pGPGSS was digested with EcoRI, and synthetic oligonucleotides MCS1F (SEQ ID NO:1) and MCS1R (SEQ ID NO:2) were annealed and then ligated into the EcoRI cut pGPGSS to produce pGPG7p (FIG. 14). In a second manipulation, primer directed mutagenesis (Stratagene La Jolla CA; QuikChange XL) using primers BglKF (SEQ ID NO:3) and BglKR (SEQ ID NO:4) was performed according to the vendor's procedures to remove the Bgl II site from the gentamycin resistance sequence producing pGPG7. A further derivative with an alternative multicloning site was made was by cutting pGPG7 with EcoRI and AscI

and ligating in annealed CC_UPPER (SEQ ID NO:5) and CC_LOWER (SEQ ID NO:6) to produce pGPG8 (FIG. 15).

6.1.2 PRODUCTION OF DONOR VECTORS: CLONING SUBTILISIN SEQUENCES INTO pGPG

A variety of donor vectors were generated by cloning subtilisin sequences from various species into the MCS of pGPG plasmids. Construction of two representative examples of such subtilisin donor vectors is described herein. In addition to the two representative examples described in detail herein, a number of other subtilisin sequences from *B. subtilis* and *B. lichenformis* strains were also successfully cloned into pGPG plasmids using completely analogous procedures.

Six hundred base pair fragments encoding the catalytic and substrate-bindings portions of subtilisins were PCR amplified from the strains 3A13 (*B. subtilis* variety amylosacchariticus) and 5A20 (*B. licheniformis*) using two internal primers (upper - SEQ ID NO:7 and lower - SEQ ID NO:8). The PCR products were cloned into a pGEM derivative using the pGEM easy T vector (Promega; Madison, Wisconsin), which employs a T/A (Clark, 1988, Nucl. Acids Res. 16:9677-86) cloning strategy, according to the vendor's protocols. The *B. subtilis* clone was digested with Eco RI and the resulting fragment subcloned into the EcoRI sites of pGPG7 to produce pGPG7-3A13. The *lichenformis* clone was digested with Spe I and Sph I and the resulting fragment subcloned into the Xba I and Sph I sites of pGPG7 to produce pGPG7-5A20. The DNA sequences of the *lichenformis* and *subtilis* inserts (SEQ ID NOs:19 and 21, respectively) were determined by standard procedures and are shown in FIG. 16. The two clones encode protein fragments (SEQ ID NOs:20 and 22, respectively) with 8 and 13 amino acid differences relative to the corresponding 200 amino-acid sequenced coding regions of the respective *lichenformis* and *subtilis* subtilisin target sequences described below.

6.2 TARGET VECTORS

6.2.1 PRE-TARGET VECTORS

Construction of pre-target vectors capable of driving expression of subtilisin sequences was performed described herein. The vectors are termed pre-target vectors because no negatively selectable marker had yet been introduced into the target sequences present on the vector. Target vector construction (whereby the negatively selectable marker is introduced into the target sequences) is described in the following section.

The vectors described in this section were constructed as derivatives of the vector pWH1520 (MoBiTec Gmbh, Götingen, Germany). pWH1520 provides selection in both *E. coli* (ampicillin resistance) and *B. subtilis* (tetracycline resistance) as well as

separate replication origins that function in these bacteria. In addition pWH1520 provides a xylose-regulated promoter (Rygus and Hillen, 1991, Microbiol. Biotechnol. 35:594-599) that is expressed in *B. subtilis*. To verify that subtilisin proteases can be expressed in this system and thereby provide an expressible target for DGA with both the *subtilis* and *lichenformis* subtilisins, intact complete *lichenformis* and *subtilis* protease coding sequences were PCR cloned from *lichenformis* (ATCC No. 14580, ATCC Manassas, Virginia) and *B. subtilis* (3A1; BGSC Department of Biochemistry, The Ohio State University Columbus, Ohio). Subtilisin from *lichenformis* was cloned using *B. lichenformis* Subtilisin forward and reverse primers (SEQ ID NOS:9 and 10) and subtilisin from *B. subtilis* was cloned using *B. subtilis* forward and reverse primers (SEQ ID NOS:11 and 12) using standard PCR conditions. Both set of primers contain appropriately oriented Kpn I and Bgl II sites, allowing the direct cloning of the PCR products as transcriptional fusions into Kpn I / Bgl II cut pW1520. Clones were first verified in *E. coli* by restriction analysis and the coding sequences of both genes were then determined by standard DNA sequencing procedures. The sequences of the *B. lichenformis* gene and seconded protein (SEQ ID NOs:13 and 14) and *B. subtilin* subtilisin genes and encoded proteins (SEQ ID NOs:15 and 16) demonstrated minor variations from those published in GenBank (see FIG. 17).

The functional nature of these clones was assessed by transformation (tetracycline at 15 µg/ml selected) into the subtilisin-defective *B. subtilis* host 1A751 (Apr-, Npr-; BGSC Department of Biochemistry, The Ohio State University Columbus, Ohio). Both plasmids promoted robust clearing zones on standard casein-agar plates (Maerki *et al.*, 1984, J. Chromatogr. 283:406-411) when supplemented with 2% xylose. In the absence of xylose the *B. subtilis* clone (pWHsub) produced no zone while the *lichenformis* clone (pWHlic) produced a reduced zone of clearing, indicating a substantial level of constitutive expression. The control plasmid pWH1520 (no insert) failed to demonstrate any zone with (2%) or without xylose.

6.2.2 SELECTABLE MARKER MODULES

A cassette containing the negatively selectable galactokinase (GalK) gene and positively selectable *aadA* gene conferring spectinomycin resistance was generated (Gal-Spec cassette; Section 6.2.3, below) for incorporation into a target vector. With respect to GalK, the GalK gene is a negatively selectable marker because, in strains with a defect in both the galactose kinase gene (galK) and a defect in the galactose epimerase gene (galE), expression of the GalK gene in the presence of galactose is lethal. When GalK is present in a target gene, therefore, selection for growth in the presence of galactose, represents selection for recombination within the target recombination module that effects loss of GalK. A cassette containng the negatively selectable sucrase gene and selectable *npt*

1 conferring kanamycin resistance was also incorporated into a target vectors, as described in Section 6.2.5, below.

6.2.3 GAL-SPEC

The Gal-Spec cassette was constructed in the vector pMOD (Epicentre; Madison, WI) that contains a multi-cloning site (MCS) between inverted19-bp repeats from the Tn5 transposon. A galactokinase-containing fragment was PCR isolated from the plasmid pKG1800 (Menzel and Gellert, 1987, J. Bacteriol. 169(3):1272-78) using the following: an upper primer (SEQ ID NO:17) and a lower primer (SEQ ID NO:18). This PCR product was digested with BglII to produce a fragment ready for cloning. Digestion of the plasmid pHP45 omega (Fellay *et al.*, 1987, Gene 52:147-54) with Bam HI and gel purification of the *aadA* harboring 2028 base pair fragment provides DNA containing the *aadA* gene which confers spectinomycin resistance. The cassette was produced by simultaneously ligating the Bam HI cut pMOD, the BglII flanked galactokinase-containing PCR product and the Bam HI bracketed *aadA* gel purified fragment. Clones were isolated by selecting for spectinomycin resistance (50 μg/ml) on Luria agar by standard techniques. Clones containing the galactokinase gene were identified by their ability to confer on a galK- host strain the ability to ferment galactose as visualized by their red color on galactose MaConkey agar (Becton Dickson, Difco Division, Franklin Lakes, NJ). One such isolate (pMODGALSPEC) was further characterized by restriction analysis to determine the relative orientation of the cloned pieces. The resultant Gal-Spec cassette is given in FIG. 18. The 4.5 Kb Gal-Spec cassette-containing Pvu II fragment from pMODGALSPEC was been used successfully for construction of target vectors by introduction of the cassette into target sequences. The target sequence insertion method utilized herein was *in vitro* transposition is described in the following section.

6.2.4 PRODUCTION OF TARGET VECTORS: TRANSPOSITION OF GAL-SPEC CASSETTE INTO TARGET SEQUENCES

Insertions into the target gene encoding the *B. subtilis* subtilisin were made into a pGPG6 derivative carrying the *B. subtilis* subtilisin *apr* gene. This derivative was made by first cloning the gene into a pGEM derivative using a T/A (Clark, 1988, Nucl. Acids Res. 16:9677-86) cloning strategy (Promega; Madison Wisconsin, pGEM easy T vector) according to the vendor's protocols following PCR amplification from the strain 1A685 (BGSC Department of Biochemistry, The Ohio State University, Columbus, Ohio) using the *B. subtilis* subtilisin Forward and Reverse Primers (SEQ ID NOS. 11 and 12). This product was then (re)cloned as an EcoR I fragment into pGPG6 using standard molecular biology techniques to produce pGPG6-sub.

To perform the transposition, the 4.5 KB Pvu II fragment with the Gal-Spec cassette flanked by the inverted 19 base pair repeats of Tn5 was purified (from the plasmid pMODGALSPEC; see Section 6.2.3 above) and mixed with equal molar quantities of pGPG6-sub (paragraph above) in the presence of transposase according to the vendors (EZ::TN transposase kit; Epicentre; Madison, WI) directions. The resultant mixture was electroporated into OTG24 and then plated on Luria plates selecting spectinomycin (50 μg/ml) according to standard procedures.

Plasmid DNAs from spectinomycin resistant (and gentamycin 10 μg/ml; pGPG6 marker) isolates represent target vectors, *i.e.*, vectors comprising target sequences into which the selectable marker cassette (which includes a negatively selectable marker) has been inserted. Target vector sequences were screened for the approximate location of the cassette insert by restriction analysis. Those located within the central 600bp region of interest (see Section 6.1.2) were sequenced to determine the precise location of the inserts. Among those two, GS10 and GS2, were subsequently used in the DGA process, as described below. FIG. 20 shows the plasmid pGPG6-sub with position of the inserts indicated. GS10 and GS2 were used in the DGA allele re-assortment process following DGA mediated transfer to the target plasmid pWHsub (see 6.5.3; below).

It is noted that while these plasmids are, indeed, target vectors, as the term is described herein, the plasmids can also be utilized as donor vectors. For example, once the selectable marker cassette is introduced into a position of interest, DGA procedures can transfer the portion of the target gene carrying the marker cassette of interest to a homologous target gene sequence present on a target vector by using the vector above as the donor vector in the DGA process.

6.2.5 PRODUCTION OF TARGET VECTORS: DIRECT CLONING OF SELECTABLE MARKER CASSETTES INTO TARGET SEQUENCES

Described in this section is the construction of target vectors by insertion of a selectable marker cassette into a target gene sequence via direct cloning methods.

To allow the direct cloning of selectable marker cassettes into target DNA sequences of pWHLic and pWHSub, extraneous sequences were deleted from the vectors to reduce them from 9 KB to approximately 3.8 KB in size. This reduction in plasmid size establishes a number of restriction enzymes sites within the target gene sequences as unique sites in these derivatives, thus allowing the direct cloning of the selectable marker cassette (including a negatively selectable marker) and otherwise facilitates their manipulation.

Deletion was accomplished by restriction enzyme-based deletion of the *B. subtilis* selectable (tetracycline resistance) marker and the *B. subtilis* replication origin.

pWHLic and pWHSub were digested (separately) with Spe I and Aat II, filling-in with T4 DNA polymerase and subsequent DNA ligation was used to re-circularize the molecule according standard procedures. The resulting vectors, pLIBsub and pLIBlic, were confirmed by a series of restriction nuclease digests and are shown in FIG. 19.

pLIBLic has unique Nde I and BsrG I sites in the central subtilisin region of interest. To produce suitable target vectors in the *lichenformis* gene the plasmid pRL250 was cut with BamH I to produce a 2.3 KB fragment containing the npt I (kanamycin resistance) and sacB (sucrase; sucrose sensitivity) cassette (Kan-Suc). The nucleotide extensions on this fragment were filled-in using T4 DNA polymerase and then ligated into Nde I or BsrG I (separately)-digested pLIBLic preparations, which had been similarly filled in. The resultant ligation mixtures were transformed into OTG 197 selecting kanamycin resistance (40 μg/ml on Luria agar). The structure of the resultant plasmids, pLIBLic-Nde and pLIB-Lic-BsrG, was confirmed by restriction analysis. pLIBLic is illustrated in FIG. 19 with the unique Nde I or BsrG I shown.

6.3 STRAINS FOR THE GROWTH AND MANIPULATION OF pGPG-DERIVED TARGET AND DONOR VECTORS

Table 1 below describes bacterial strains employed in the generation and use of target and donor vectors derived from pGPG plasmids.

TABLE 1

Strain	Genotype
OTG 2	ΔlacX74 galE galK thi rpsL ΔphoA
OTG 24	DE3(lac) uidA (ΔMluI)::pir(wt)
OTG 27	endA hsdR pro supF / pRK2013::Tn9
OTG 82	ΔlacX74 galE galK thi rpsL ΔphoA *mut*L218::Tn10
OTG 83	ΔlacX74 galE galK thi rpsL ΔphoA zei::Tn10
OTG 197	DE3(lac) uidA (ΔMluI)::pir(wt) / pRK2013 ::Tn9

The galactose resistance selection requires a strain with a defect in both the galactose kinase gene (galK) and a defect in the galactose epimerase gene (galE). In such a strain, expression of GalK from the Gal-Spec cassette (described in Section 6.2.3) is lethal in the presence of galactose and selection for growth in the presence of galactose is a selection for loss of the cassette. The bacterial strain OTG2 (also known as KS272; Dr.

Stanley Maloy, http://salmonella.life.uiuc.edu/strainfinder.html) has defective galK and galE genes.

The genetic background of OTG2 was modified to include a tetracycline resistance element suitable for selection against donor strains that do not have the tetracycline resistance element. To accomplish this, bacteriophage P1 was grown on RFM 101 (Menzel and Gellert, 1987, Proc. Nat'l Acad. Sci. U.S.A. 84(12):4185-9; zei::Tn10) and CSG 7050 (Singer *et al.*, 1989, Microbiol Rev 53(1):1-24; mutL218::Tn10) and used to transduce OTG 2 to growth on tetracycline- containing media (Luria Agar plus 25 μg/ml tetracycline) to produce OTG82 and 83, respectively. The *mut*L218 of OTG 82 abolishes mismatch repair. Due to the loss of mismatch repair function, the rate variant production via recombination between less homologous sequences using the procedures described herein is increased.

The use of the pGP704 derivatives as a donor of DNA requires transacting π replication functions, and for conjugative transfer, a mobilizing element. Strains supporting the growth of the plasmids and directing conjugal transfer are well known. Among these are strains OTG 24 (see Metcalf *et al.*, 1994, Gene 138:1-7) and OTG 27 (see Ely B. 1985 Mol Gen Genet 200:302-4). The strain OTG197 was constructed from these strains by conjugal transfer of pRK2013 ::Tn9 (from OTG 27) into OTG24 on minimal media plates containing 40 mg/ml chloramphenicol. OTG197 was used in all mating experiments below to transfer donor vectors into target cells for variant formation (see Section 6.4).

6.4 CO-INTEGRANT FORMATION

The experiments described in this section demonstrate successful use of the first step of a two-step variant selection using the DGA methods of the invention. For ease of discussion, this first step is referred to herein as co-integrant formation. Use of the term, however, as discussed above, is not intended to bind the subject matter of the invention to a particular theory or mechanism.

In the crosses described below (Table 2), two different target cell strains were used: OTG82 (ΔlacX74 galE galK thi rpsL ΔphoA mutL::TN10) and OTG83 (ΔlacX74 galE galK thi rpsL ΔphoA zei::Tn10) to host the target vectors. Both target strains were transformed with the negatively selectable target plasmid pWHsub-GS2 that was formed by DGA recombination (see Section 6.5.3 below). A set of donor plasmids containing the DNA encoding the central 200 amino acids of the *apr* gene from different wild type variants (cloned into the EcoR I site of pGPG7; see Section 6.1.2) were used in the crosses. To form co-integrants donor strains with the designated pGPG7 derivatives in the genetic background of OTG197 were grown selectively (plus gentamycin 10 μg/ml, 40 μg/ml chloramphenicol) overnight from isolated single colonies in liquid Luria broth.

Target strains were grown selectively in Luria Broth with ampicillin (100 μg/ml) in OTG82 or OTG83.

To perform the crosses, 5 microliters of donor were spotted on the surface of a Luria broth plate together with 5 microliters of target. After the 10 microliter spot dried into the plate (10-30 minutes), the mating mixtures were transferred to an incubator at 37°C for 4-6 hours. At the end of this incubation interval the patch was transferred with a sterile applicator stick to 200 microliters of Luria broth in the well of a microtiter plate. This 200 microliter aliquot was thoroughly mixed to resuspend the cells and 10 microliters were spotted and spread on a Luria broth plate with 10 μg/ml gentamycin (to select for the pGPG replicon) and 15 μg/ml tetracycline (to select for against the pGPG harboring host strain derived from OTG97). These plates were incubated overnight (14-16 hours) and the number of colonies growing from the various crosses and control (donor and target alone) were scored. The results are tabulated in Table 2:

TABLE 2

Donor Sequence	Colonies Mut+ Target	Colonies Mut- Target	Colonies Donor Alone	A.A. Differences Relative to Target
3A1	58	124	0	none
3A3	112	237	0	1
3A6	132	215	0	1
3A7	4	17	0	24
3A11	14	212	0	5
3A13	4	48	0	15
3A14	2	18	0	25
None (pGPG7)	2	12	0	N/A

The results show that co-integrant formation in the Mut-plus strain was significantly above background when five or fewer differences exist (out of 200) amino acids. In the Mut-defective background this was extended to 15 differences. For the various strains listed, the nucleotide changes noted were approximately 3 times those seen at the amino acid level as numerous silent mutations were seen. The placement of a large insert with DGA (see Section 6.5.3, which describes a donor sequence containing a selectable marker cassette) demonstrated that large segments with no homology can be recombined into foreign sequences provided sufficient flanking homology exists.

Eight colonies from each of the crosses with the Mut-defective host were selectively purified by isolating single colonies of agar media (10 μg/ml gentamycin and 15 μg/ml tetracycline) for subsequent Gal-resistance mediated co-integrant resolution (Section 6.5.1). Following purification, the clones were grown (selectively) in liquid and then frozen with 10% glycerol as a cryo-preservative. Such frozen cultures can be used as a source of resolvable co-integrants at a later date. Failure to purify and grow selectively leads to large-scale segregation (>50% gentamycin negative) of the donor plasmid sequences.

Results from a second set of co-integrant forming crosses are shown in Table 3 below. The targets in this set are the Kan-Sac insertions (pLIBLic-Nde and pLIB-Lic-BsrG) described above (Section 6.2.5) transformed into the, OTG82 mutS::TN10 host, and results were identical with both inserts. Donors were clones of the core 200 amino acid encoding sequence from a set of various wild type *lichenformis* subtilisins as described in Section 6.1.2. Procedures employed were identical to those described above for the *B. subtilis* subtilisin donors and pWHsub-GS2.

TABLE 3

Donor Sequence	Colonies Mut-Target	Colonies Donor Alone	A.A. Differences Relative to Target
5A2	>500	0	9
5A20	>500	0	8
5A30	>500	0	7
5A36	>500	0	0
(none) pGPG7	20	0	N/A

These results are consistent with those in Table 2. The presence of homology dramatically stimulated the formation of gentamycin resistant colonies. Based on the numbers above >95% of the gentamycin resistant colonies observed could be attributed to the presence of a shared regions of homology with typical wild type variant sequences. Small microscopic background gentamycin resistant colonies appeared on all plates, which can be attributed to spontaneous events occurring in the target strain as these colonies were also seen in target alone controls. Such colonies were readily distinguishable from the large true co-integrants.

6.5 CO-INTEGRANT RESOLUTION

The experiments described in this section demonstrate successful use of the second step of a two-step variant selection using the DGA methods of the invention. In

particular, following section describes phenotypic selection for the DGA-directed resolution of co-integrant based on the lethality of galactose (Section 6.5.1), as described in Section 5.1.2.1.1. For ease of discussion, this second step is referred to herein as co-integrant resolution. Use of the term, however, as discussed above, is not intended to bind the subject matter of the invention to a particular theory or mechanism.

6.5.1 GAL-BASED RESOLUTION

Eight co-integrants each from the crosses summarized in Table 2 were streaked for single colonies (from cultures cryo-preserved in 10% glycerol; described in Section 6.4)) on Luria broth plates with ampicillin (100 μg/ml), spectinomycin (50 μg/ml) and gentamycin (10 μg/ml). Single colonies were inoculated into Luria broth liquid (without drug) and incubated overnight at 37°C with gyratory shaking. Ten microliters (each) from these cultures were spread on a MacConkey Agar (base; Becton Dickson, Difco Division, Franklin Lakes, NJ) plated with 2% galactose and 100 μg/ml ampicillin. Following overnight growth three types of galactose-resistant colonies appeared on the agar surface: red colonies (with various morphologies), white opaque colonies and white translucent colonies, in numbers varying from a few dozen to several hundred. Resolved co-integrants were among the white translucent colonies, and a single white colony was picked from each spot, and re-streaked for purification on the same MacConkey agar. These purified white colonies were tested for spectinomycin resistance (50 μg/ml) and gentamycin resistance (10 μg/ml). The resultant colony types are summarized below in Table 4 (Gent = gentamycin, Spec = spectinomycin, R = resistance, and S=sensitivity)

TABLE 4

Original Donor Sequence	GentS, SpecS	GentR, SpecS	GentS, SpecR	GentR, SpecR
3A1	8/8			
3A3	6/8		2/8	
3A6	4/8		2/8	2/8
3A7			3/8	5/8
3A11	5/8		1/8	2/8
3A13	2/8		5/8	1/8
3A14			5/8	3/8
None (pGPG6)			7/8	1/8

The phenotype consistent with the co-integrant resolving recombination is "GentS, SpecS." Such a phenotype indicates that the sequence that includes Spec is lost, as are the sequences associated with the Gent-conferring donor vector. Subsequent plasmid purification and restriction analysis demonstrated that this class of galactose resistant colony had a gross structure identical with that of pWHsub, demonstrating loss of the negatively selectable Gal-Spec insert.

Transformation of these plasmids into *B. subtilis* host 1A751 (double protease defect) demonstrated that they all produce active protease. DNA sequence analysis showed that they have, in most instances, inherited alleles from the donor plasmid. To further analyze the uptake of sequences from donor vectors, the two co-integrants from the 3A13 cross which gave rise to the GentS, SpecS galactose-resistant colonies were re-plated and eight new GentS, SpecS galactose resistant colonies were isolated from each for DNA sequence analysis (see Section 6.6.2 below).

6.5.2 MOLECULAR SELECTION

The following sections describe the use of molecular methods for selecting for variant target molecules. Section 6.5.2.1 describes digesting a population of DNA molecules subjected to DGA with an enzyme whose restriction site is present in the original target vector but absent from the variant target molecule produced by DGA. Thus, unrecombined target vectors are digested by the restriction enzyme and, because they are not linear, are not take up by new host cells, while variant molecules are not linearized by the enzyme and can be selected for by transformation and growth on selective media. Section 6.5.2.2 describes the use of PCR to identify variant target molecules that have lost negatively selectable marker sequences in the selection process.

6.5.2.1 RESTRICTION ENZYME-BASED RESOLUTION

The Kan-Suc insert in pWLIB-Lic-BsrG contains a unique Xho I restriction site not present in pWHLIB-Lic. According to the strategy above such a site should work to select DGA-directed recombinant molecules (strategy diagramed in FIG. 21). Co-integrants of pGPG7-5A20 (Section 6.1.2) were formed with pWLIB-Lic-BsrG (Section 6.2.5) according procedures described above for the *B. subtilis* subtilisin crosses (Section 6.4). A collection of approximately 500 gentamycin and tetracycline resistant colonies was pooled and plasmid DNA was prepared according to standard procedures. This DNA was digested overnight with excess Xho I according to the vendor's recommendations (New England BioLabs, Beverly, MA). The Xho I-digested DNA preparation was then further treated with phosphatase according to standard procedures and used to transform OTG82 selecting ampicillin (100 µg/ml) resistance on Luria Broth agar plates with 5% sucrose. Twenty-six

of these colonies were further tested for gentamycin resistance (10 μg/ml; to test for the presence of donor sequences) and kanamycin resistance (40 μg/ml; to test for loss of the insert). Seventeen of the twenty-six had the correct phenotype and were digested with Kpn I and BamH I to test for the presence the apr sequences (less the insert). All clones proved to be correct. The central 600 base pair region of the subtilisin gene was sequenced in these recombinant clones. The results are shown below in Section 6.6.1, and demonstrate that several variant subtilisin coding regions were generated, each of which encoded a variant subtilisin polypeptide exhibiting protease activity.

6.5.2.2 PCR-BASED SELECTION

To test the PCR selection strategy, co-integrants were formed as described in Section 6.4 (Table 2) with pGPG7 donor plasmids with the 3A1, 3A7, 3A11 gene sequences and pGPG7 alone. Gentamycin and Tetracycline selected colonies from these crosses were pooled (about 500 colonies each, separately) and DNA prepared according to standard procedures. This DNA, along with control DNA from pWHSub, was used as substrates for PCR reactions (29 cycles; 1 min. 93°C, 1.5 min. 57°C, 1.5 min. 72°C) employing the primers originally described for the isolation of the *B. subtilis* subtilisin coding sequences (see Section 6.2) Products from the PCR reaction were resolved using agar gel electrophoresis with a 0.8% gel employing standard conditions.

The gel-resolved products from this experiment and the strategy for the PCR selection are shown together in FIG. 22. The gel revealed that a PCR product with a size appropriate to the *B. subtilis* subtilisin coding sequences was seen for the unit length gene (pWHsub) but not the gene containing the insert pWHSub-GS2. The unit length product is also noted for pools of DNA derived from the co-integrants made from the 3A1 and 3A11 pGPG7 donors. Co-integrant resolution experiments based on phenotypic selection (galactose resistance) above (Table 4) show that properly resolved structures were readily isolated from 3A1 and 3A11.

6.5.3 DGA-BASED SEQUENCE INSERTION

Section 6.2.4 describes the isolation of Gal-Spec cassettes in the donor plasmid pGPG7-sub, giving rise to plasmids GS2 and GS10. Using such an insert-containing sequence in a donor vector allows the sequence containing the insert to be moved (repeatedly, if desired) into target vectors using DGA. This, therefore, represents an efficient way to create new target recombination modules.

A culture with the pGPG7sub-GS2 plasmid (host strain OTG24) was grown and mixed with a second culture containing the target pWHSub (host strain OTG83) and co-integrants were selected (gentamycin and tetracycline) as described above (Section 6.4).

Individual colonies were purified and the co-integrant structure was confirmed by noting the unselected co-inheritance of spectinomycin resistance and galactose sensitivity.

Two methods were used to isolate resolved structures. In one strategy co-integrants were grown non-selectively in Luria broth and plated for single colonies on agar media containing ampicillin (100 μg/ml) and spectinomycin (50 μg/ml). Plates with isolated single colonies were replica printed to a second agar plate containing ampicillin, spectinomycin, and gentamycin (10 μg/ml). Individual colonies that were ampicillin and spectinomycin resistant but gentamycin sensitive (a marker for the donor sequence) appeared at a frequency of 0.5%. Restriction analysis of these plasmids demonstrated the desired recombinant product.

In a second strategy a restriction enzyme-based molecular selection was employed to isolate the desired recombinants. To do so DNA was prepared from a pool of co-integrants by standard procedures and digested with the restriction enzyme BsrG I which cuts in the pGPG7 sequences but does not cut in the Gal-Spec insert, the coding sequences for subtilisin or the pWH1520 vector. This digestion result in making the co-integrant linear but leaving the desired resolved structure as a circle molecule. The digestion mixture was treated with phosphatase according to standard procedures and used to transform OTG83 selecting spectinomycin (50 μg/ml) resistance. Individual colonies (6) were purified and all proved to be ampicillin and spectinomycin resistant but gentamycin sensitive. Subsequent restriction nuclease analysis showed the expected DNA structure. Phenotypic tests demonstrated the desired galactose sensitive growth. One such colony was retained and used for the crosses described above in Section 6.4. The work required and reagents used to recover the desired resolved structure was substantially less when the molecular selection was applied; 100% of the colonies had the desired structure as opposed to 0.5% in the unselected screened sample.

The movement of the insert to a target cell is illustrated in FIG. 23. These steps show how DGA (with the molecular restriction nuclease-based selection) can be used to insert donor sequences into a stretch of homologous target DNA. In the example, extensive homology extending across the entire subtilisin encoding sequences was used. This homology could have been limited to confine the extent of the subtilisin-encoding sequences participating in the event. Thus, in addition to removal of DNA sequences by DGA (*e.g.*, removal of negatively selectable markers from target vectors; see, *e.g.*, Sections 6.4 and 6.5.1, *supra*), DGA can be used to insert DNA sequences into target modules. Combining removal and insertion can be used to introduce non-homologous sequences into a target gene, as illustrated in FIG. 9. The non-homologous sequences can, *e.g.*, comprise a selectable marker, or a coding sequence intended to become part of the modified target gene.

6.6 RESULTS

Sections 6.5.1 (3A13 by 3A1) and section 6.5.2.1 (5A20 by 5A36) describe the production of recombinant molecules by a galactose-based and molecular selection-based DGA, respectively. To further investigate the nature of these recombinants, 3 milliliter samples were grown up in Luria Broth with 100 μg/ml ampicillin and plasmid DNA was prepared according to standard procedures (Qiagen; 28159 Avenue Stanford, Valencia CA). The DNA sequences of the recombinant molecules were analyzed using Vector NTI software (Informax, 7600 Wisconsin Avenue, Suite #1100, Bethesda, MD). Results from those analyses are discussed below.

6.6.1 5A20 BY 5A36 CROSSES

Of the seventeen DGA recombinants derived from the 5A20 by 5A36 cross (described in Section 6.5.2.1, *supra*), DNA sequence results were obtained for thirteen recombinants. The thirteen sequenced samples defined twelve unique molecules distinguished by re-assortments of the 30 DNA sequence differences between the 5A36 target and the 5A20 donor molecules. All re-assortments were simple rearrangements consisting of contiguous patches of 5A20 sequences replacing stretches of the 5A36 sequence as would be expected from a single double crossover event ensuing from the DGA selected recombination event. No mosaics suggesting multiple crossover events were noted. These DGA exchanges were executed in a mutL strain that precluded mismatch repair, which may give rise to apparent multiple crossover events. Above it was noted that mutL was required for effective co-integrant formation in instances of significant sequence divergence. It is possible that co-integrant structures could have been moved to a mismatch repair proficient strain where mosaics could be observed. In the absence of multiple crossover events, 465 unique molecules are possible from single crossover events between two molecules with 30 differences. The pooling of large numbers co-integrants and the subsequent molecular selection (by restriction digestion) is an effective method of obtaining a random collection of recombinant molecules, which in the instant example yielded 12 out of 13 unique sequences.

To analyze proteins produced from these molecules, the predicted protein sequences were determined by *in silico* translation (Vector NTI). The resulting coding sequences were aligned, showing that the 12 variants produced represent seven different variant proteins. That is, some DNA variants produced encode the same variant protein. These results also demonstrate that co-integrant selection leads to a family of sequence variants once the co-integrant is resolved.

6.6.2 3A13 BY 3A1 CROSSES

The sequences of fifteen galactose-resistance-selected recombinants from the 3A13 by 3A1 cross (described in Section 6.5.1 above) were obtained. To analyze the proteins produced from these molecules the predicted protein sequence was determined by *in silico* translation (Vector NTI). The results showed that seven different variant proteins were produced. As above, therefore, some of the DNA variants produced encode the same variant protein. As also shown, these results further demonstrate that co-integrant selection leads to a family of variants upon co-integrant resolution.

Finally, each of the products encoded by the sequence variants produced by DGA in both crosses (including those for which sequence was not determined) demonstrated functional protease activity by the casein-agar test following introduction into a *B. subtilis* host. DGA selection is a highly effective way to obtain novel re-assorted structures.

6.6.3 CONCLUSION

The results described herein demonstrate the successful use to DGA to generate subtilisin variants using the techniques described in Section 5.2 above. Not only was a very high yield of nucleic acid variants generated, these nucleic sequences encoded a variety of subtilisin variant polypeptides, all of which exhibited subtilisin protease activity. Thus, the present invention provides methods of generating variant polypeptides in a more directed, efficient and cost-effective manner than the presently available methods of directed evolution.

The invention described and claimed herein is not to be limited in scope by the specific embodiments herein disclosed since these embodiments are intended as illustration of several aspects of the invention. Any equivalent embodiments are intended to be within the scope of this invention. Indeed, various modifications of the invention in addition to those shown and described herein will become apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims. Throughout this application various references are cited, the contents of each of which is hereby incorporated by reference into the present application in its entirety for all purposes.